О.Ю. Колесниченко, Г.Н. Смородин, И.В. Ильин,
О.В. Журенков, Л.С. Мазелис, Д.А. Яковлева, В.Л. Дашонок

Текстовая аналитика Big Data: перспективы для суперкомпьютеров

Аннотация. В статье представлены результаты первого этапа многоцентрового исследования по аналитике Больших данных, которое организовано по инициативе Академического Партнерства EMC в России и СНГ. Процесс Data Mining был осуществлен благодаря использованию многоуровневой гибридной схемы с доступом к неклассическим суперкомпьютерам, к которым относятся Google и Яндекс. Студенты участвовавших в исследовании вузов получали практические навыки в ходе лабораторных работ, а также включали Data Mining в свои персональные дипломные работы. Основной вопрос, который ставился в данном исследовании, касался оценки того, что представляет собой Интернет с точки зрения хаотичного заполнения его огромным количеством текстовых массивов. Показано, что неструктурированные массивы ключевых слов, относящиеся к категории Big Data, отражают в информационной среде Интернета реальные процессы, происходящие в глобальном социуме. Массивы ключевых слов можно использовать для прогностической оценки состояния государств.

Ключевые слова и фразы: Большие данные, Big Data, Data Mining, текстовая аналитика, неклассические суперкомпьютеры, Google, Яндекс.

Введение

В мировом рейтинге суперкомпьютеров Топ 500 российские суперкомпьютеры представлены с 22 места (Суперкомпьютерный комплекс МГУ). Всего в списке из 500 самых мощных суперкомпьютеров мира находятся 9 российских суперкомпьютеров. Парк суперкомпьютеров в целом по России намного больше, но увеличение их численности ограничено, так как стоимость таких установок превышает несколько миллионов рублей и достигает нескольких миллиардов рублей. На фоне малодоступности суперкомпьютерных технологий сильно растет потребность в высокопроизводительной аналитике Больших данных (High Performance Data Analytics, HPDA) во всех сферах экономики. Например, Правительство Великобритании [1] определило 9 критических технологических сфер, все они требуют применения суперкомпьютерных технологий и должны приоритетно финансироваться:

1. Энергетика (Energy Storage);
2. Большие данные (Big Data);
3. Орбитальные спутники (Satellites);
4. Робототехника (Robotics and Autonomous Systems);
5. Синтетическая биология, генная инженерия (Synthetic Biology);
6. Регенеративная медицина, стволовые клетки (Regenerative Medicine);

7. Сельскохозяйственные науки (Agri-Science);
8. Улучшенные материалы (Advanced Materials);
9. Квантовые технологии (Quantum Technologies).

Все 9 направлений (помимо отдельного упоминания Big Data как самостоятельной сферы) связаны с Big Data, с HPDA, и с новым технологическим вектором – NBIC-конвергенция (природа вещества – нано, формы жизни – био, свойства разума – когно, информационный обмен – инфо). NBIC-конвергенция полностью базируется как на HPDA, так и на применении суперкомпьютеров. В документе Конгресса США [2] указано, что Национальный научный фонд (National Science Foundation) должен приоритетно инвестировать на долгосрочной основе в проекты по Data Mining и накопление Big Data, что в итоге должно сменить парадигму в научно-исследовательской среде и в сфере образования. Удовлетворять потребность в проведении научных работ, практическом обучении студентов и осуществлении задач для разных сфер бизнеса в рамках аналитики Больших данных (Big Data) позволяют многоуровневые гибридные схемы организации Data Mining.

Текстовая аналитика Big Data является одним из приоритетов в области HPDA. Так или иначе, все упирается в создание алгоритмов «понимания» текстов, их смыслового распознавания. В этом ключ к достижению конечной цели всего компьютерного программирования – создание искусственного интеллекта. Подходы к анализу текста могут быть очень разными. Так, в Исследовательском центре искусственного интеллекта Института программных систем им. А.К. Айламазяна РАН [3] выделяют следующие виды текстовой аналитики Big Data: морфологический, синтаксический и семантический. Сегодня широкое распространение получил такой вид аналитики, как Sentiment Analysis [4], его можно отнести к семантическому упрощенному анализу. Есть и варианты морфологического анализа, когда осуществляется подсчет определенных ключевых слов в зависимости от заданных условий. Например, проект Google N-Grams Corpus предоставляет такой сервис по годам встречаемости слов (Google управляет массивом слов объемом более 1 трлн из отсканированных литературных источников).

В зависимости от того, какая база неструктурированных датафицированных (оцифрованных) текстовых данных используется, определяют цели аналитики и выдвигают гипотезы. В настоящее время идет осмысление всех возможностей для текстовой аналитики, и пока ее четкой классификации нет. Может быть поставлена задача сбора информации о персонах или продукции. Или целью будет прогностический антитеррористический контроль, чем уже много лет занимается The National Security Agency в рамках программы PRISM. Помимо суперкомпьютерных задач, которые следует решать по мере развития HPDA – или еще появился термин Advanced Analytical Theory and Methods [4] – в рамках текстовой аналитики необходимо решать задачи из области лингвистики, психологии, социологии, глобалистики, политологии.

Одна из задач, которая разрабатывается еще со времен появления первых

прогностических алгоритмов-моделей, созданных профессором Массачусетского технологического института Джейем Форрестером в конце 60-х годов прошлого века [5], заключается в прогнозе развития глобальных процессов. В дальнейшем, ориентируясь на эти модели Форрестера, США стали корректировать свою внешнюю политику и влиять в заданном ключе на глобальную политику. Задача, которая появилась позже – составление рейтингов государств по результатам анализа данных из разных сфер. Наиболее устоявшиеся и популярные рейтинги: Index of Economic Freedom (The Heritage Foundation) [6], Networked Readiness Index (World Economic Forum) [7], Fragile States Index (Fund for Peace) [8]. В сфере глобальной безопасности в рамках работы научных центров НАТО проводится аналитика Больших данных из закрытых ресурсов с построением прогностических моделей устойчивости государств в аспекте развития внутригосударственной гражданской войны. Например, интересна работа канадского эксперта Paul Comeau по прогнозу стабильности государств региона MENA (Middle East & North Africa) [9].

Ранее было показано, что глобальное управление требует становления целостности всей совокупности международных институтов [10]. Для этого нужна тотальная алгоритмизация, которая будет основана на применении суперкомпьютеров. Только такой подход, базирующийся на HPDA, с использованием глобального доступа к открытым данным, позволит со временем унифицировать методы, что и приведет к формированию целостности системы глобального управления. Идеальный вариант подобной прогностической аналитики – HPDA в режиме реального времени. Рейтинги государств будут влиять на стратегии бизнеса по инвестированию, колебания курса валют, решения по военному вторжению. На какой вид аналитики будет опираться подобная рейтингующая алгоритмизация? Определенно можно сказать, что текстовая аналитика займет одно из важнейших мест. Мировое сообщество пройдет определенный путь эволюции алгоритмизации – будет большое разнообразие предлагаемых алгоритмов, затем наступит период выработки общемировых подходов к алгоритмизации, и будут выбраны некоторые из них в качестве узаконенных международным правом инструментов глобального управления.

В данном исследовании в центре внимания – использование неструктурированных текстовых Больших данных (Big Data) для поиска подходов в прогностической оценке внутренней политической и экономической ситуации в государствах. Важно понять, что представляет собой информационная среда Интернета (текстовый компонент), как она характеризует государства, и насколько неструктурированные текстовые массивы, накапливаемые в Интернете хаотично, из разных источников и по разным поводам, коррелируют с классической статистической информацией. Задача сложная и большая, требующая не одного исследования, а пошагового анализа с усложнением подходов. В данной статье обсуждается первый шаг на этом пути, примененный подход к текстовой аналитике – морфологический, не

семантический. Ключевые слова собирались из открытых Интернет-ресурсов: Web pages, Social Network (эта категория ресурсов относится к неструктурированным Большим данным [4]). Не было никакого поиска графов (онтологических графов) и оценки смысла текстов. Задача ставилась оценить некоторые корреляции ключевых слов Big Data со статистическими показателями.

Ранее морфологический подход к оценке неструктурированных текстовых массивов, накапливающихся в Интернете, был применен в проекте «Google Flu», компания Google отслеживает динамику появления в Интернете определенных ключевых слов (в основном это запросы, вводимые людьми в окно поиска). Специалисты Google определили 45 условий поисковых запросов, которые имеют высокий коэффициент корреляции с официальными эпидемиологическими статистическими данными по заболеваемости гриппом [11, 12]. В результате удается фактически в реальном времени получать информацию о том, в каких регионах начинается эпидемия гриппа. Этот опыт аналитики Больших данных прежде всего ценен тем, что была установлена связь между хаотичным появлением в Интернете конкретных ключевых слов и реальными событиями, происходящими в социуме. С момента появления проекта «Google Flu» к неструктурированным текстовым массивам Интернета стали относиться, как к данным, из которых может быть извлечена важная информация.

При таком подходе к аналитике огромных массивов ключевых слов акцент делается на базовую характеристику Big Data – объем (Volume). До появления Интернета человечество не имело такого опыта, как работа со словами в качестве данных, тексты печатались на бумажных носителях, и не было возможности объединить воедино то, что печатали (или писали) люди (не только статьи, но также и телеграммы, письма, которые сегодня во многом заменили соцсети). Теперь же тексты легко объединить в один массив при проведении Data Mining, а также диапазон напечатанных и оцифрованных слов для анализа сильно расширился: живая разговорная речь, телефонные разговоры, телеграммы и письма (отражающие оперативную обстановку) стали заменяться социальными сетями, блогами, комментариями под статьями и под новостями, и поисковыми запросами. Интернет способствует тому, что речевая деятельность человека все больше переходит в письменную оцифрованную форму, то есть Интернет датафицирует речевую функцию человека. Значит, если люди пишут, думают и говорят о чем-то больше или меньше, то это должно иметь какие-то причины. Слова стали данными, анализируя которые можно получить какое-то знание о текущей обстановке или сделать выводы о векторе развития ситуации.

Неструктурированные текстовые массивы Больших данных в Интернете относятся к той категории данных, которые подходят для быстрого мониторинга ситуации. Подчеркнем, что в исследовании не рассматривается так называемый Sentiment analysis, подход принципиально иной. Интернет рассматривается как

неструктурированная среда, постоянно заполняемая печатными словами. Интенсивность появления тех или иных слов имеет динамику. В случае поиска отражения в Интернете имиджа государств, динамика накопления определенных слов может характеризовать обстановку в государствах. Если говорить о более сложном и требующем большей мощности анализе, то логично начать с работы по установлению онтологических графов и онтологической среды для каждого из анализируемого языка (в данном исследовании анализировались слова на русском и английском языках). В HPDA (или Advanced Analytical Theory and Methods) выделяют метод Graph Theory. Этот метод можно обозначить как следующий шаг к семантическому анализу текстов. Считают, что компьютерное понимание текста достигается за счет трех действий: 1) погружение текста в единую среду знаний – онтологию, 2) формальное представление смысла в памяти компьютера и 3) возможность операций над онтологическим смыслом [13]. В объеме Big Data такие задачи можно ставить только перед очень мощными суперкомпьютерами. Но, для начала нужны этапы предварительной работы, установление графов, обнаружение кластеров и построение единой среды знаний. Для этапа аналитики по методу Graph Theory те ключевые слова, которые в данном исследовании показали сильные корреляционные связи со статическими показателями, могут быть взяты в качестве категорий-узлов онтологических графов.

Помимо интереса к построению рейтингов государств, исследование имеет особую актуальность в связи с тем, что глобализация способствует применению ранее не существовавших методов воздействия на социум именно в сфере мягкой силы (Soft Power). Как утверждает профессор Факультета глобальных процессов МГУ им. М.В. Ломоносова О.Г. Леонова [14, 15], среди деструктивных методов или глобальных политических технологий воздействия особое место занимают мягкие глобальные политические технологии и социокультурное воздействие (включая технологии управления массовым сознанием). Применение таких методов делает людей мишенью в глобальной перспективе, подрывая основу для устойчивого глобального развития. Либо, социальные процессы могут возникать спонтанно и индуцироваться какими-либо факторами, в том числе новыми техническими трендами. Все это требует разработки методов оценки, измерения, прогнозирования, а также выявления катализирующих социальные процессы факторов.

1. Постановка задач и методика

Можно обозначить основной вопрос, который ставился в данном исследовании: оценка того, что представляет собой Интернет с точки зрения хаотичного заполнения его огромным числом разных слов, есть ли в таком подходе какая-то логика и связь с реальными событиями в глобальном социуме?

Исследование является уникальным проектом по аналитике Больших

данных (Big Data Analytics), которое впервые организовано в России по инициативе Академического Партнерства EMC в России и СНГ. Академическое Партнерство EMC служит открытой площадкой для всех заинтересованных в ИТ-тематике вузов России и региона СНГ, а также входит в более широкую сеть глобального Академического Партнерства корпорации EMC (EMC Academic Alliance), объединяющую вузы США и других стран мира. Одной из важных задач Академического Партнерства EMC является формирование будущего рынка труда и создание прослойки высококвалифицированной рабочей силы, ориентированной на новые тенденции и технологии [16].

В данной работе представлены результаты первого этапа многоцентрового исследования, в котором приняли участие 3 вуза, входящие в Академическое Партнерство EMC: Московский Государственный Университет им. М.В. Ломоносова, Алтайская Академия экономики и права и Владивостокский Государственный Университет экономики и сервиса (рисунок 1).

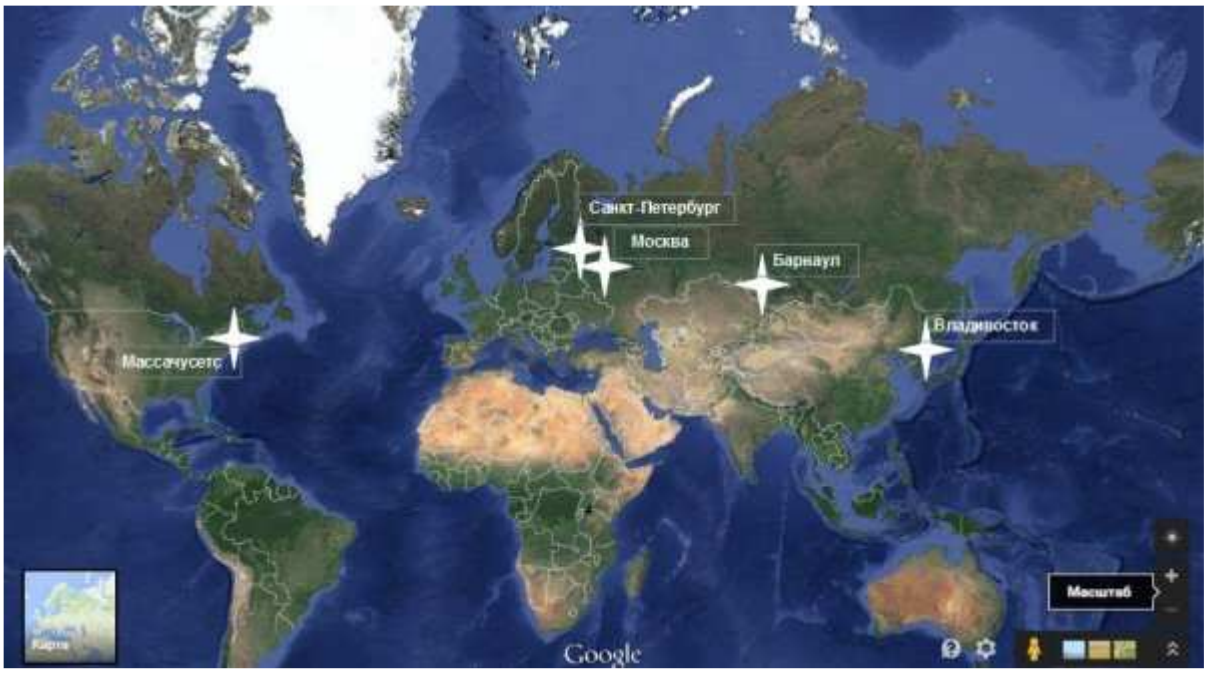


Рис. 1. Карта многоцентрового исследования Big Data Академического Партнерства EMC.

Пилотное название данного исследования – «Третья Волна», по аналогии с термином, введенным американским политологом и философом Элвином Тоффлером [17]. «Третья Волна» по Тоффлеру – это современная информационная эра, меняющая мир и способствующая процессам глобализации.

Таблица 1.
Многоуровневая гибридная схема организации Data Mining
в многоцентровом исследовании «Третья волна»

High Performance Computing Case-study Big Data
Microsoft Excel: F-тест, коэффициент корреляции Пирсона, дополнительная визуализация
R-STUDIO | Google+ API | Яндекс API Поиска по блогам
Datafication Codebook, визуализация. Распределенные и параллельные вычисления. Internet surface, Cloud Technology
GitHub | Microsoft SQL 2012 Enterprise Edition
20 (Intel Core i5-2300, SDRAM 4GB) + 7 (Intel Core i5-650, SDRAM 2GB) = 27 CPU = 94 ядра, 94 GB RAM | Серверное оборудование: AMD 6386SE, AMD 6282SE, Intel Xeon E5-2690 v2, Intel Xeon E5-2690, Intel Xeon X5660, Intel Xeon E5-2667 v2, 16 процессоров, 18 тыс. 432 ядер. Дисковые ресурсы, установленные непосредственно в серверное оборудование: 2.5" Intel SSD S3700 400ГБ, 3.5" Seagate 3ТБ SATA 7.2K, 2.5" Seagate 300ГБ SAS 10000, 2.5" Seagate 147ГБ SAS 15000, 3.5" Seagate 300ГБ SAS 15000 Системы хранения данных: IBM DS3524 Express Dual Controller 2ГБ, IBM EXP3524 Duial IO module Ленточные автозагрузчики: IBM TS2900
Неклассические гибридные суперкомпьютеры: **Яндекс** – 10 дата-центров, MapReduce; **Google** – 14 дата-центров, MapReduce Web pages, Social Network

Основное положение, которое было сформулировано на этапе формирования плана исследования – *проводится анализ открытых Интернет-ресурсов (неструктурированных массивов ключевых слов), которые представляют собой информационное, сгенерированное людьми, **отражение** реальных политических, экономических и социальных процессов, а не эти процессы как таковые.* Согласно этому основному положению ставились задачи исследования и рассматривались полученные результаты.

Команда исследователей состояла из двух групп: специалисты в области ИТ (математики), которые осуществляли по заданиям Data Mining (изъятие необходимых данных большого объема из открытых источников Интернета и представление их в виде таблиц и графиков), и гуманитарии, те кто формулировал задания и затем анализировал полученные данные. Учитывая большой объем работы, в процессе Data Mining участвовали студенты старших курсов под руководством своих научных руководителей. Использовались поисковые системы Google и Яндекс, которые считают неклассическими гибридными суперкомпьютерами [18, 19]. Многоуровневая гибридная схема организации процесса Data Mining представлена в таблице 1.

В охват попадали все доступные для этих поисковых систем открытые текстовые ресурсы: блоги, социальные сети, микроблоги, а также новостные публикации и всевозможные статьи и комментарии. В данном случае был применен главный принцип Data Mining для Big Data – сбор неструктурированной разнородной информации без каких-либо урезающих фильтров. Получаемые датафицированные показатели варьировали по численности в основном от нескольких сотен до нескольких десятков миллионов. Выгруженные данные обрабатывались методом гиперкуба, с использованием языка R.

ИТ-специалисты, выполняя данную работу, ставили перед собой дополнительные математические задачи по разработке новых ИТ-подходов и ИТ-инструментов для текстовой аналитики Big Data. Студенты вузов получали практические навыки в ходе лабораторных работ, а также включали Data Mining в свои персональные дипломные работы. На всех этапах выполнения Data Mining участники со стороны команды ИТ-специалистов получали необходимую консультативную помощь от Академического Партнерства EMC.

Описание полученных результатов Data Mining – таблиц и графиков – выполняла группа гуманитариев. Они же составляли первичное задание по Data Mining, которое представляло собой протокол единой формы, четко описывающий все условия процесса Data Mining. Этапы получения результатов были следующими:

- Case-study Big Data – разработка гипотезы по оценке отражения в информационной среде имиджа государств и их экономических показателей;
- Datafication Codebook – детальный перечень всех необходимых характеристик для извлечения данных (текстовые ключевые слова и фразы, а также название стран и годы, в которые публиковались тексты, на английском и

русском языках);

- Сетка датафикации – детальное описание мэшапа (Mash-up, сопоставления разных блоков Больших данных), сопоставление представляло собой следующую формулу: характеристика + (возможно, еще характеристика) + год публикации + название страны;
- Первичная визуализация – по результатам мэшапа группой ИТ-специалистов составлялись графики;
- Вторичная визуализация – дополнительная графическая обработка результатов осуществлялась группой гуманитариев на основе предоставленных им табличных результатов, в целях выявления скрытых трендов и тенденций и описания результатов;
- Аналитика Big Data – описательный анализ результатов Data Mining, который включал дополнительную статистическую обработку конечных табличных данных и сопоставление их с другими исследованиями в анализируемой области.

2. Имиджевые группы: позитивные и негативные

Для осуществления ранжирования государств в зависимости от образа (имиджа), который складывается в процессе появления упоминаний их в глобальной сети Интернет, было отобрано несколько характеристик – ключевых слов в привязке к периоду от 2000 года до 2015 года. Ключевые слова следующие: терроризм (terrorism), террорист (terrorist), оккупация (occupation), наркотики (narcotic), насилие (violation), демократия (democracy), развитие (development). В таблице 2 показан список 17 стран, вошедших в исследование, и соответствующая каждой стране доминирующая по количеству упоминаний характеристика из перечисленного выше списка ключевых слов.

Можно выделить группу государств, для которых в информационном поле Интернета из анализируемого списка характеристик в преобладающем большинстве случаев встречались негативные, связанные с темой терроризма. Вторая группа государств обсуждается в привязке к понятию «развитие», что можно отнести к позитивному информационному фону в отношении этих государств. Также можно выделить два случая с негативным контекстом, привязанным к теме наркотиков, и два случая с позитивным контекстом, привязанным к теме демократии. При этом следует отметить, что данное распределение по группам отражает *преобладающий информационный фон в отношении этих государств, «измеренный» путем оценки встречаемости заранее выбранных ключевых слов-характеристик (массивов слов)*. Понятия «позитивная» и «негативная» группы нельзя воспринимать с точки зрения стабильной и благополучной обстановки в стране. Речь идет о «зеркальном» отражении имиджа страны в глобальном информационном поле.

Таблица 2.
Принадлежность анализируемых государств
к позитивной или негативной группе по имиджу

Ключевое слово (характеристика), Data Mining в привязке ко всему периоду с 2000 года по 2015 год							
Страна	Терроризм	Террорист	Оккупация	Наркотики	Насилие	Демократия	Развитие
Азербайджан							Позитивная
Афганистан				Негативная			
Грузия							Позитивная
Израиль		Негативная					
Ирак		Негативная					
Иран							Позитивная
Йемен		Негативная					
Киргизия							Позитивная
Китай							Позитивная
Ливия						Позитивная	
Пакистан		Негативная					
Палестина	Негативная						
Сирия		Негативная					
Тунис						Позитивная	
Турция							Позитивная
Узбекистан				Негативная			
Украина		Негативная					

Так, Ливия (позитивная группа, «Демократия») перенесла войну, свержение и публичное убийство Муаммара Каддафи, но тема демократии стала лидирующим трендом в глобальном информационном фоне в отношении этой страны. Тунис (позитивная группа, «Демократия») является «иконой» Арабской Весны – революционной волны борьбы за демократию, охватившей арабские страны региона MENA (Middle East & North Africa), см. таблицу 3. Именно с 18 декабря 2010 года, после самосожжения бедного продавца Мохаммеда Буазизи, начались массовые протесты в Тунисе, а затем и в других арабских странах.

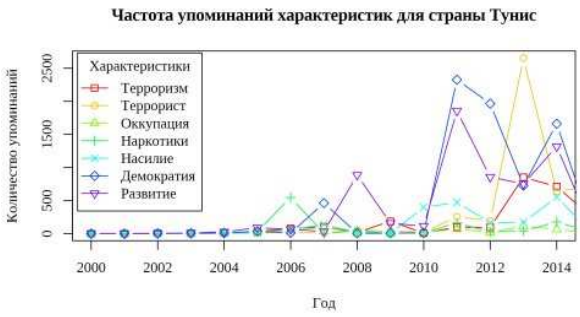
Все страны разные по процессам, происходящим в них, и обсуждение их в Интернете также неоднородно. Более того, графические данные Data Mining показали уникальный профиль для каждой из стран в отдельности. Но, задача исследователей в области Big Data и заключается в поиске общих закономерностей в массивах очень разнородных данных. Например, характеристика «демократия» может упоминаться и в негативном контексте, как отсутствие демократии. Но, в этом исследовании применялись подходы к

анализу Big Data, где неточность и неоднородность, отсутствие фильтрации – являются важным условием исследования [11, 12]. Для уточнения общей картины просматривались такие характеристики, как «насилие», «терроризм», «оккупация». На примере Ливии видно, что все эти негативные признаки, указывающие на обсуждение отсутствия демократии, встречаются намного реже, чем характеристика «демократия». В статье приведены графики по данным Data Mining в качестве примеров (рисунок 2).

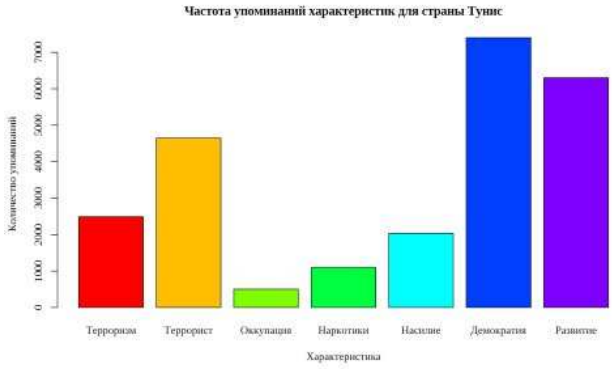
Таблица 3.
Хронология начала Арабской Весны
в некоторых странах региона MENA
(https://ru.wikipedia.org/)

Тунис	17 декабря 2010 года
Ливия	13 января 2011 года
Йемен	18 января 2011 года
Сирия	26 января 2011 года
Ирак	10 февраля 2011 года

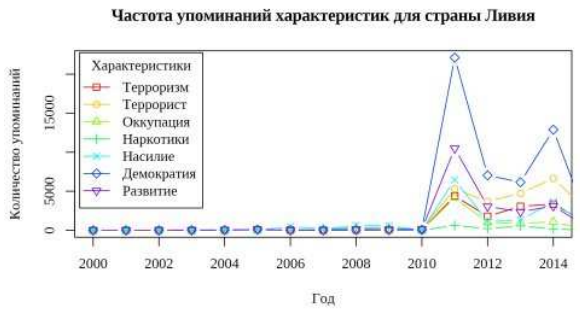
2-A



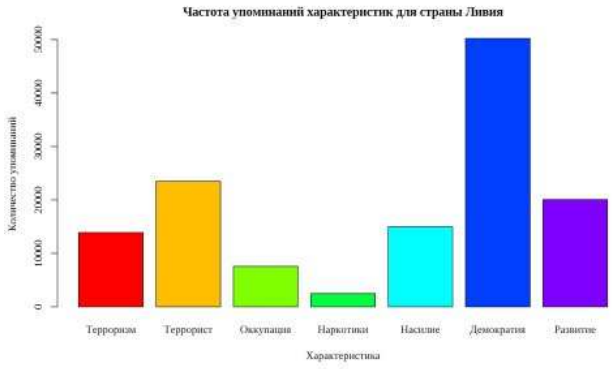
2-B



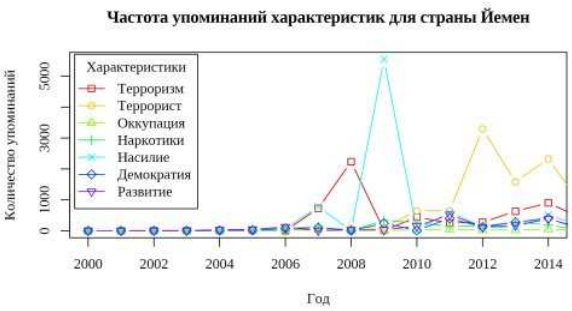
2-C



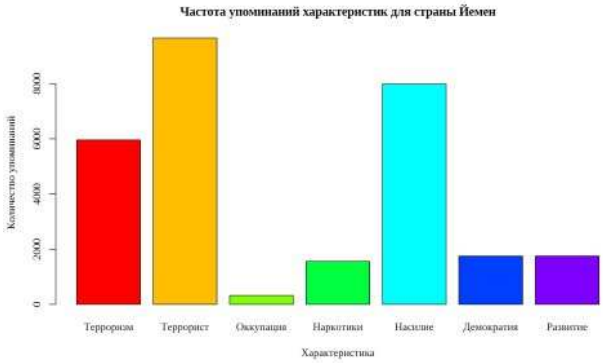
2-D



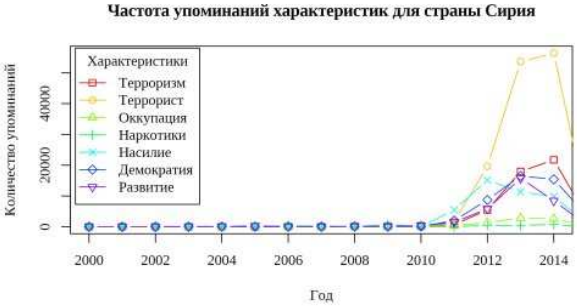
2-E



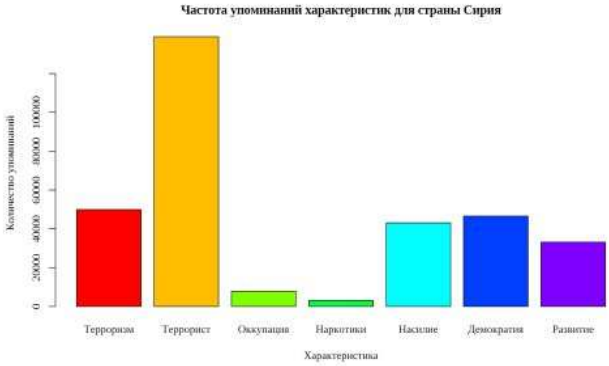
2-F



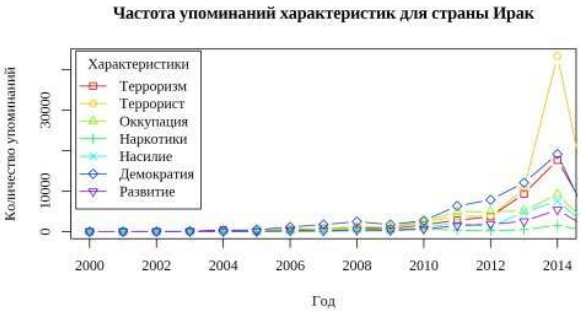
2-G



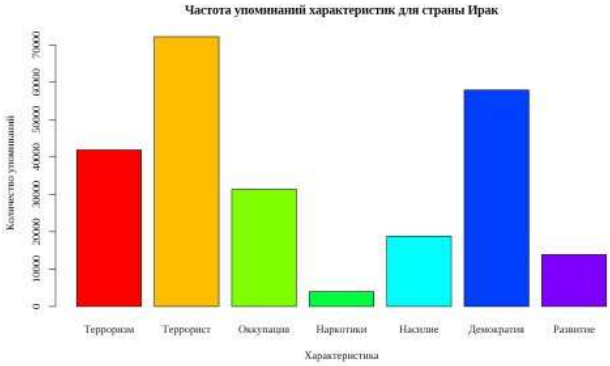
2-H



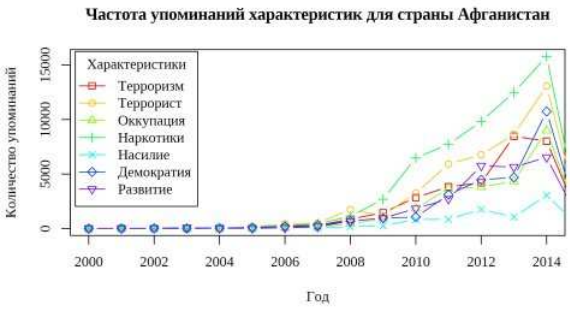
2-I



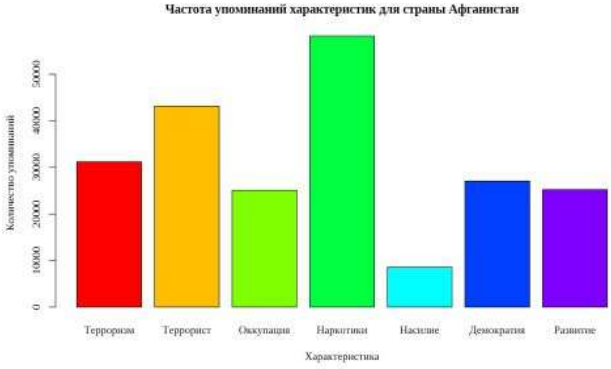
2-J



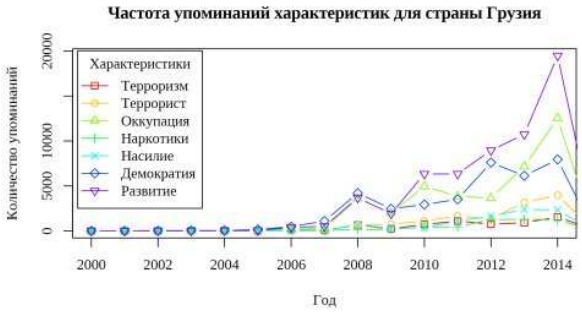
2-K



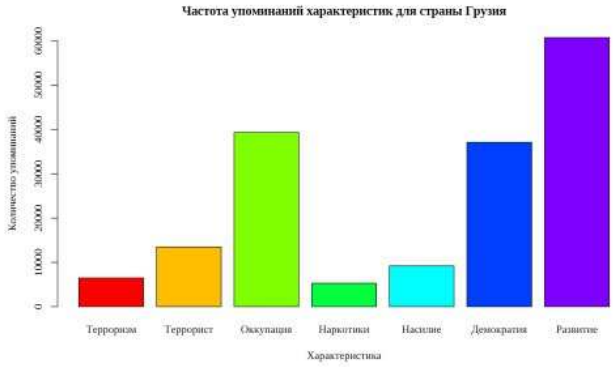
2-L



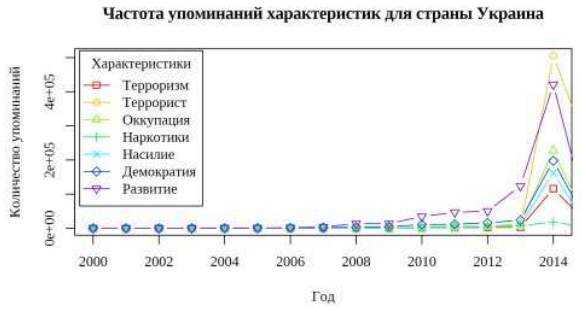
2-M



2-N



2-O



2-P

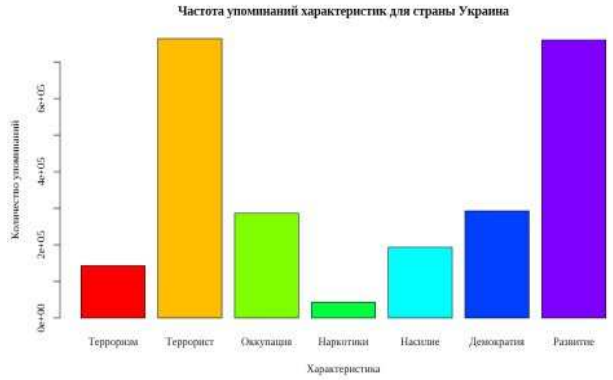


Рис. 2. Примеры графиков по результатам Data Mining в привязке к периоду с 2000 года по 2015 год. A, B – Тунис (позитивная группа, «Демократия»); C, D – Ливия (позитивная группа, «Демократия»); E, F – Йемен (негативная группа «Террорист»); G, H – Сирия (негативная группа «Террорист»); I, J – Ирак (негативная группа «Террорист»); K, L – Афганистан (негативная группа «Наркотики»); M, N – Грузия (позитивная группа, «Развитие»); O, P – Украина (негативная группа «Террорист»).

3. Сравнительный анализ двух потоков Data Mining

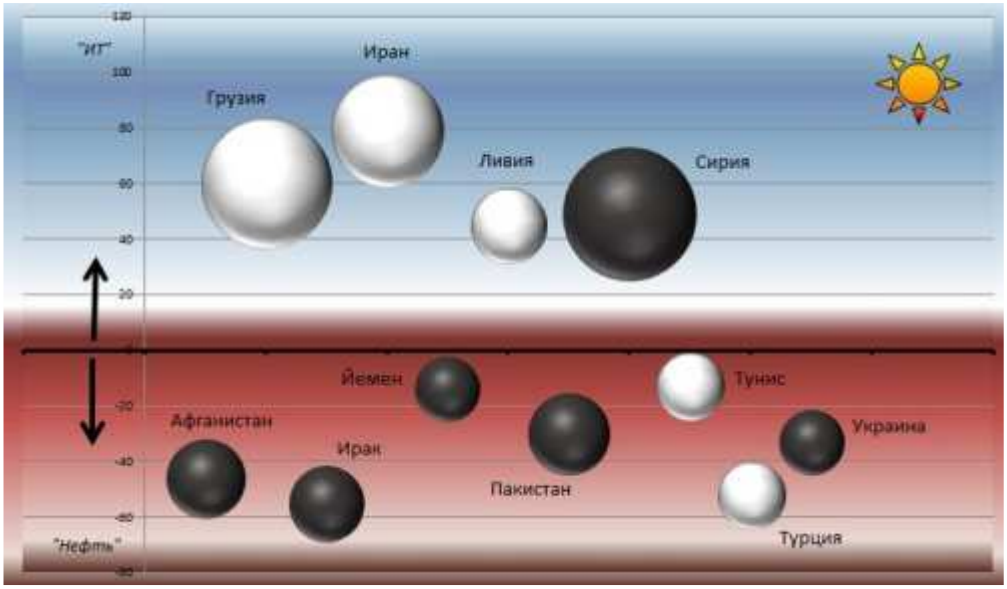
Для 11 государств, ранжированных по имиджевым группам, был сделан анализ упоминаний ключевых слов, указывающих на направления развития экономики. Ключевые слова анализировались в привязке к двум годам их публикации – кризисный 2008 год и 2015 год. Ключевые слова были следующими: капельное орошение (drip irrigation), мобильный телефон (mobile phone), солнечные батареи (solar panel), атомная электростанция (nuclear power plant), нефть (oil). Учитывая то, что 2008 год вошел в анализ полностью, а 2015 год лишь частично (первый квартал), то при сравнительном анализе внимание обращалось только на положительный прирост частоты встречаемости

характеристики, что точно подтверждало тенденцию более частого ее упоминания в сравнении с 2008 годом. Из 11 государств 5 относятся к позитивному по имиджу информационному фону («Развитие» и «Демократия»), а 6 – к негативному по имиджу фону («Террорист» и «Наркотики»).

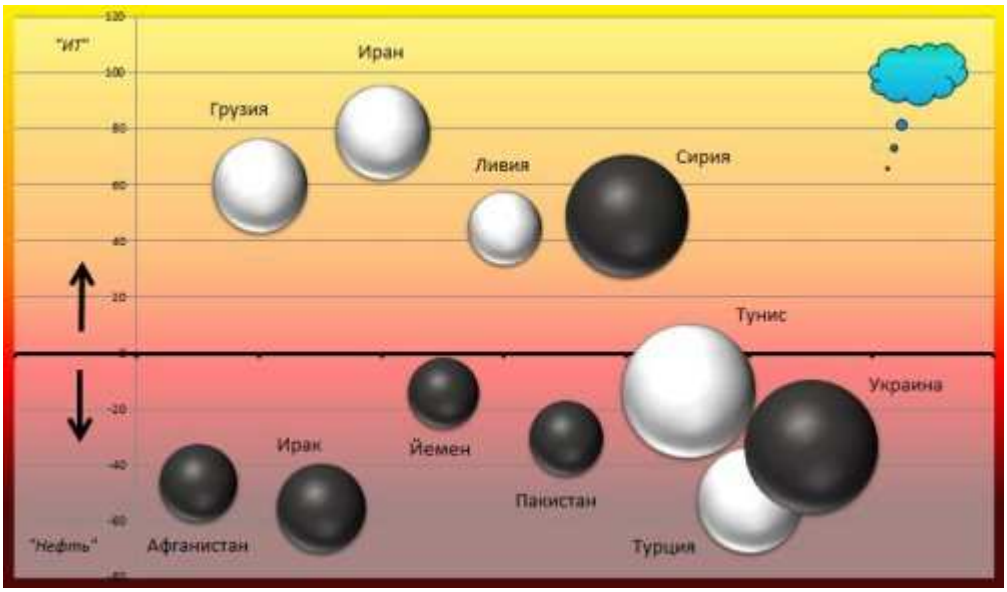
В таблице 4 сведены вместе данные от двух групп ИТ-специалистов, выполнявших Data Mining по разным заданиям, что повышает значимость результатов сравнения, так как не было никаких субъективных факторов, которые могли бы повлиять на общую картину. И также, благодаря организации исследования как многоцентрового, привнесен требуемый элемент разнообразия в аналитику Big Data.

Все 11 стран, взятые для сравнительного анализа, в кризисный 2008 год характеризовались информационным фоном, отражавшим сырьевой уклад экономики – группа «Нефть» (лидировала по встречаемости характеристика «нефть»). При этом, к 2015 году для 4-х стран информационный фон поменялся – лидирующей характеристикой стали ключевые слова «мобильный телефон», что позволило отнести эти страны в группу «ИТ» (информационные технологии). Стоит отметить, что 3 страны из группы «ИТ» относятся к позитивным группам «Развитие» и «Демократия». Четвертая страна – Сирия, которую сравнительный анализ позволил охарактеризовать как страну, имеющую хорошие тенденции, указывающие на потенциал развития, но попавшую в негативную группу из-за гражданской войны и террористических событий, происходящих на ее территории. Данный вывод помогла сделать дополнительная визуализация результатов с построением пузырьковой диаграммы (рисунок 3). После выявления неоднородности позитивных и негативных групп, страны были разделены на типичных (зерновых) представителей группы и нетипичных (попавших в группу из-за влияния каких-то факторов). В таблице 4 жирным шрифтом обозначены страны, являющиеся типичными (зерновыми) представителями позитивной или негативной группы, а нежирным шрифтом – нетипичные. На рисунке 4 представлены графики, показывающие динамику доминирующего тренда по экономическим характеристикам.

Для группы стран, имеющих смену доминирующего тренда с 2008 года по 2015 год «Нефть / ИТ», среднее арифметическое по приросту в % частоты встречаемости характеристики «солнечные батареи» составило 223±137%, что на 89% выше среднего арифметического данного показателя для группы стран с неизменившимся доминирующим трендом «Нефть / Нефть», составившего 134±85% (различия недостоверны). Среднее арифметическое по этому же показателю («солнечные батареи»), но для негативной по имиджу группы стран («Террорист», «Наркотики») составило 155±102%, что на 16% ниже среднего арифметического данного показателя для позитивной группы стран («Развитие», «Демократия»), составившего 171±120% (различия недостоверны).



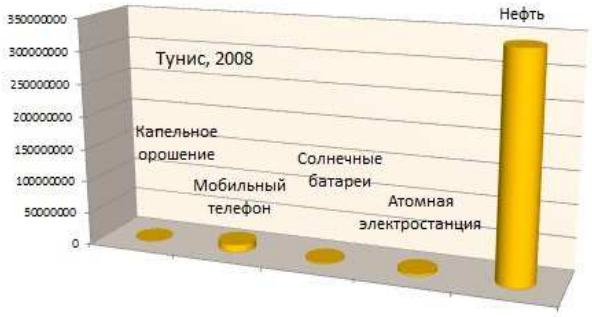
3-A



3-B

Рис. 3. Взаимосвязь характеристик по Data Mining, 2015 год. Размер круга – количество упоминаний характеристик «солнечные батареи» в млн. (A) и «капельное орошение» в млн. (B). По оси Y – количество упоминаний характеристик «мобильный телефон» и «нефть» в млн. Светлые круги – позитивная группа стран, темные круги – негативная группа стран.

4-A



4-B

4-C

4-D

4-E

4-F

4-G

4-H

4-I

4-J

Рис. 4. Графики по результатам Data Mining в привязке к 2008 г. и 2015 г.: A, B – Тунис; C, D – Ливия; E, F – Йемен; G, H – Сирия; I, J – Ирак.

Таблица 4.
Динамика анализируемых экономических характеристик
в позитивной и негативной группах стран по имиджу

Страна	**Группа по имиджу**	**Динамика встречаемости характеристики, прирост в % с 2008 года к 2015 году**					**Группа по экономи-ческому тренду, 2008 / 2015**
		капельное орошение	мобильный телефон	солнечные батареи	атомная электро-станция	нефть	
Афганистан	«Наркотики»	-	-	139,00%	7,00%	-	Нефть / Нефть
Грузия	«Развитие»	34,00%	-	381,00%	-	-	Нефть / ИТ
Ирак	«Террорист»	-	-	109,00%	-	-	Нефть / Нефть
Иран	«Развитие»	42,00%	362,00%	136,00%	-	-	Нефть / ИТ
Йемен	«Террорист»	-	-	-	-	-	Нефть / Нефть
Ливия	«Демократия»	-	315,00%	152,00%	-	-	Нефть / ИТ
Пакистан	«Террорист»	-	-	303,00%	-	-	Нефть / Нефть
Сирия	«Террорист»	8,00%	128,00%	-	-	-	Нефть / ИТ
Тунис	«Демократия»	345,00%	-	89,00%	-	-	Нефть / Нефть
Турция	«Развитие»	49,00%	-	97,00%	-	-	Нефть / Нефть
Украина	«Террорист»	9,00%	-	70,00%	-	-	Нефть / Нефть

Интересно то, что эти данные иллюстрируют имеющуюся тенденцию [20], когда в рамках темы устойчивого глобального развития и построения будущего обязательно говорят о солнечной энергетике, в то время как тема нефти зачастую связана с негативными чертами современного мира – война, терроризм, борьба за ископаемые энергоресурсы. По полученным данным в результате Data Mining можно определенно заметить признаки происходящей в настоящее время смены технологического уклада с постепенным переходом на

информационные технологии и солнечную энергетику. При этом заметно, как данный *энергетический переход позитивно влияет на имидж государств*. В целом, информационные технологии все больше втягивают в себя все секторы экономики через тотальную датафикацию (с постепенным формированием Internet of Things, Интернета вещей). Этот процесс и является главным триггером смены технологического уклада – массовое применение сенсоров и различных датчиков, передающих информацию в Интернет, требует автономного энергопитания во все больше и больше возрастающих объемах.

На рисунке 3 видно, что страны, относящиеся к позитивной группе по имиджу («Развитие» и «Демократия», обозначены светлыми кругами), характеризуются более частным упоминанием таких характеристик, как «солнечные батареи» и «капельное орошение» (количество встречаемости характеристик в млн). Эти же «позитивные» государства расположены в зоне «ИТ» (доминирующий тренд в 2015 году в ряду выбранных для данного исследования характеристик). А государства, относящиеся к негативной группе по имиджу («Террорист», обозначены темными кругами), характеризуются меньшим упоминанием характеристик «солнечные батареи» и «капельное орошение», и они расположены в зоне «Нефть» (доминирующий тренд в 2015 году в ряду выбранных для данного исследования характеристик).

Данное сопоставление двух независимых потоков Data Mining позволило определить государства, которые можно назвать нетипичными в своей группе. Сирия, входящая в негативную группу по имиджу, в остальном имеет принципиально положительные позиции. Это может свидетельствовать об искусственности вызванного в стране кризиса, о его индуцировании извне государства. А вот Ливия, несмотря на принадлежность к группе позитивного имиджа, имеет слабые позиции, так как тренды, указывающие на развитие, для этой страны выражены слабо и больше характерны для негативной группы. Турция и Тунис, относящиеся к позитивной группе по имиджу, также имеют слабые позиции, что может обусловить для них плохой прогноз.

4. Корреляционный анализ Big Data и классических статистических данных

Было выявлено, что датафицированная характеристика «терроризм» имеет сильную отрицательную корреляционную связь (r= -0,75) со статистическим показателем ВВП на душу населения (GDP per Capita PPP, $, база данных Index of Economic Freedom [6]. Это свидетельствует в пользу того, что неструктурированные массивы слов, в частности, слово «терроризм» отражает реальные процессы в обществе, связанные с распространением терроризма, который, как известно, связан с низким уровнем доходов населения. Чем беднее население страны, тем оно более подвержено вовлечению в террористическую активность. И с другой стороны, чем беднее государство, тем менее оно способно защитить себя от повторяющихся террористических актов. Еще одним

доказательством того, что характеристика «терроризм» отражает реальные процессы, явилось обнаружение достоверного отличия (F-тест, $p<0,05$) при сравнении значений Index of Economic Freedom в позитивной и негативной по имиджу группах государств. В позитивной группе (Азербайджан, Грузия, Иран, Киргизия, Китай, Ливия, Тунис, Турция) IEF составил 58,7±9,7; в негативной группе (Афганистан, Израиль, Ирак, Йемен, Пакистан, Палестина, Сирия, Узбекистан, Украина) IEF составил 54,7±9,7. Эти результаты показывают, что страны, определенные в негативную по имиджу группу согласно проводимому анализу характеристик Big Data, имеют более низкий показатель индекса IEF, то есть в этих странах экономические условия хуже, чем в странах позитивной по имиджу группы.

Выявлена сильная корреляционная связь между датафицированными характеристиками Big Data «демократия» и «мобильный телефон» ($r>0,7$). Характеристика Big Data «мобильный телефон» отражает в информационном поле Интернета важные социальные процессы в глобальном обществе, на что указывает обнаруженная сильная корреляционная связь с характеристикой «демократия», а с характеристиками «террорист», «терроризм», «насилие» были выявлены корреляционные связи средней силы. При этом, не было выявлено корреляции между частотой встречаемости характеристики Big Data «мобильный телефон» и статистическим показателем «количество абонентов мобильной связи» (Mobile phone subscriptions/100 pop, база данных Networked Readiness Index / World Economic Forum [7], по архиву ежегодных публикаций The Global Information Technology Report). В связи с этим можно говорить о том, что характеристика Big Data «мобильный телефон» в данном исследовании не связана со статистическим увеличением числа пользователей мобильных телефонов в странах.

Были обнаружены корреляции статистического показателя «количество абонентов мобильной связи» (Mobile phone subscriptions/100 pop) сразу с несколькими датафицированными характеристиками Big Data: «террорист», «терроризм», «насилие», «демократия», при этом сильные положительные корреляции относились к 2011 году. Такое совпадение навело на мысль о влиянии на корреляцию Арабской Весны. Для подтверждения данного предположения был проведен более подробный корреляционный анализ, охватывающий перечисленные 4 характеристики Big Data и динамику статистического показателя «количество абонентов мобильной связи» за период с 2009 года по 2015 год по архиву ежегодных публикаций The Global Information Technology Report [7]. В результате предположение подтвердилось, при этом следует отметить, что в связи с нестабильной обстановкой в арабских странах, включенных в анализ, не все погодовые публикации The Global Information Technology Report содержали информацию о статистических показателях той или иной страны. Однако проведение корреляционного анализа в расширенной группе стран (11 стран: Азербайджан, Афганистан, Грузия, Ирак, Йемен, Китай, Ливия, Сирия, Тунис, Турция, Украина), а не только среди арабских стран,

позволило допустить некоторую мозаичность и единичное выпадение в тот или иной год статистических показателей по той или иной арабской стране. В общем, для всех 11 стран, за единичным исключением в разные годы, имелись статистические данные по количеству абонентов мобильной связи.

Рассмотрев поверхностные диаграммы, построенные как корреляционное поле, можно говорить о выявлении пространственно-временной структуры (рисунок 5) как отражения в Интернете политического явления «Арабская Весна», во взаимозависимости с распространением и использованием населением мобильных телефонов (условно по внешнему виду напоминает образ «улитки», что можно обозначить как Snail-структура). Сильные положительные корреляционные связи обнаружены в привязке к 2011 году для характеристик «террорист» (terrorist), «терроризм» (terrorism), «насилие» (violation), «демократия» (democracy). Пространственно-временная Snail-структура схожа для всех 4-х анализируемых характеристик. Находка исследования указывает на то, что массивы ключевых слов, накапливающиеся в Интернете, могут стать датафицированными, измеряемыми показателями (с использованием визуализации), указывающими на реальные процессы, происходящие в глобальном обществе. Данная пространственно-временная структура в рамках заданных параметров исследования отражает сильную корреляционную связь между вышеописанными 4-мя датафицированными характеристиками Big Data, привязанными к 2011 году, и распространением мобильных телефонов, начиная с 2011 года и далее. Можно сделать вывод, что насыщение региона персональной мобильной связью произошло к 2011 году, что и стало одним из катализаторов массовых волнений арабского населения. Также обнаружена корреляционная связь, повторяющаяся для всех 4-х датафицированных характеристик Big Data, привязанных к 2014-2015 годам, с уровнем распространения мобильных телефонов в 2009 году. Объяснить такую ретроградную корреляционную связь представляется затруднительным. Опираясь на полученные результаты, требуется продолжение исследования в данном направлении.

Получив корреляции, относящиеся к статистическому показателю «количество абонентов мобильной связи» (Mobile phone subscriptions/100 pop), был проведен расширенный корреляционный анализ с такими статистическими показателями, как «число пользователей Интернета» (Individuals using Internet, %) и «вовлеченность населения в социальные Интернет-сети» (Use of virtual social networks), база данных Networked Readiness Index / World Economic Forum, по архиву ежегодных публикаций The Global Information Technology Report. Ни одной сильной корреляционной связи (то есть достигнувшей значения 0,7 и выше) не было выявлено, поэтому рассматривать эти поверхностные диаграммы с точки зрения их структуры не имеет смысла. Социальные Интернет-сети в данном исследовании не коррелируют с политическими характеристиками, что свидетельствует о менее значимой роли этого Интернет-сегмента в политических процессах в обсуждаемой группе

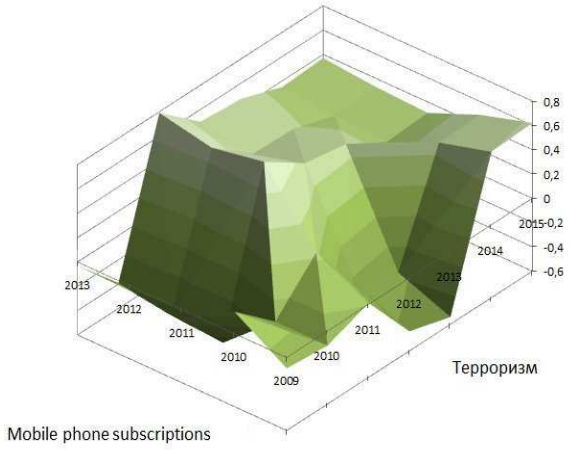
стран.

Результаты данного исследования были сопоставлены с результатами исследования Paul Comeau (2012 год) из канадского агентства Defence Research and Development Canada [9], характер которого авторами был определен как Data Analytics: в нем в сложные математические алгоритмы включались статистические показатели из открытых баз данных и закрытых оборонных баз данных. Одним из компонентов анализа был FSI – Fragile States Index, публикуемый ежегодно вашингтонским исследовательским центром Fund for Peace [8]. В работе канадских исследователей был описан прогноз политической нестабильности для государств MENA (Forecasting Collective Political Violence), при этом была выделена группа государств с высоким риском политических волнений (таблица 5).

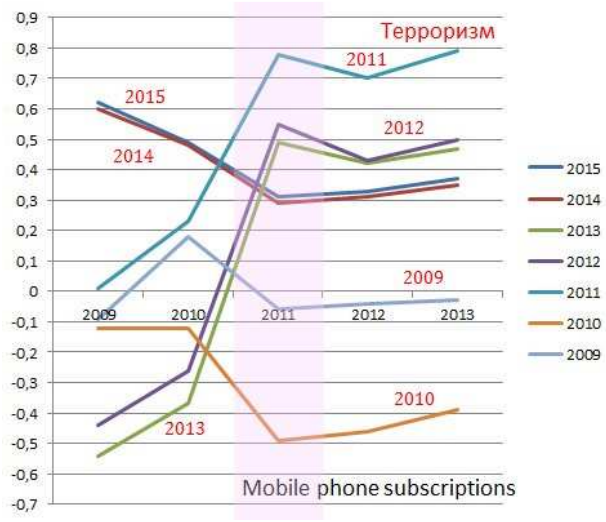
Таблица 5.
Сравнение государств, подвергшихся Арабской Весне,
по группам, определенным по результатам Data Mining

Страна	**Группа по имиджу**	**Группа по экономическому тренду, 2008 / 2015**	**Прирост характеристики «мобильный телефон» в % с 2008 года к 2015 году**	**Fragile States Index 2014**
Тунис	«Демократия»	Нефть / Нефть	Прирост отсутствует	High Warning, Rank 78
Ливия	«Демократия»	Нефть / ИТ	315,00%	Very High Warning, Rank 41
Йемен	«Террорист»	Нефть / Нефть	Прирост отсутствует	High Alert, Rank 8
Сирия	«Террорист»	Нефть / ИТ	128,00%	High Alert, Rank 15
Ирак	«Террорист»	Нефть / Нефть	Прирост отсутствует	High Alert, Rank 13

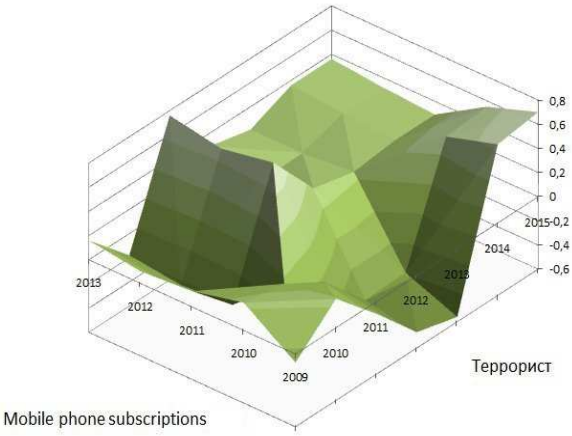
Все эти государства находятся в той части рейтингового списка, где расположены нестабильные страны по FSI-2014. Из них Тунис и Ливия более приближены к стабильной части рейтинга (которая начинается с рейтинговой позиции 139), и по данным нашего исследования Тунис и Ливия имеют позитивный имидж в Интернете по ключевым словам. Йемен, Сирия и Ирак находятся в начале рейтинга нестабильных стран по FSI, и по данным нашего исследования они имеют негативный имидж.



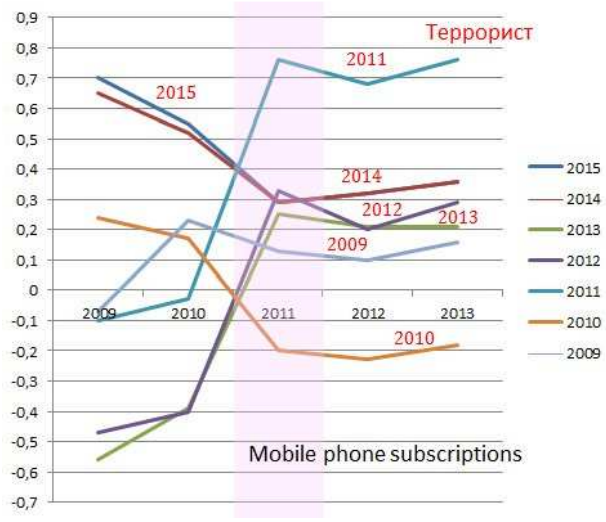
5-A



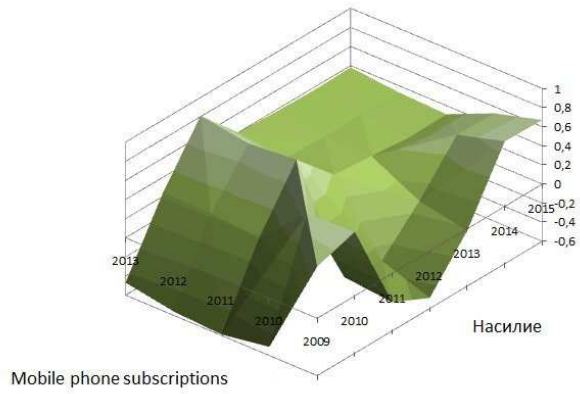
5-B



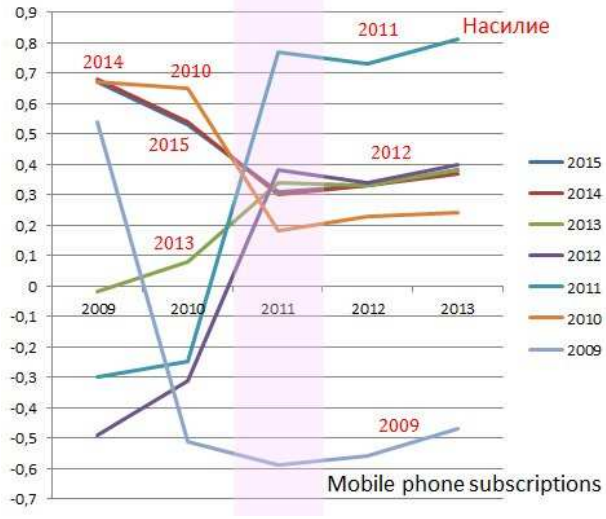
5-C



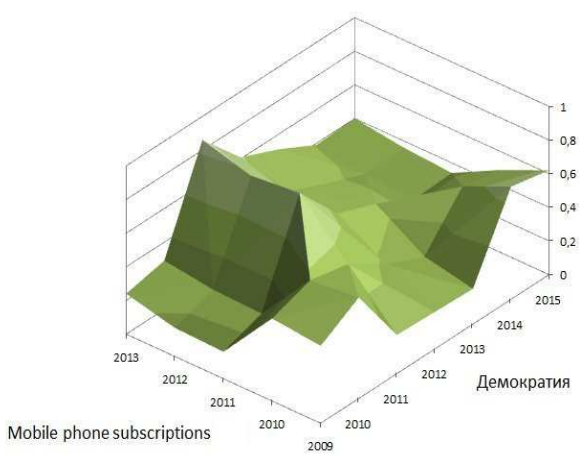
5-D



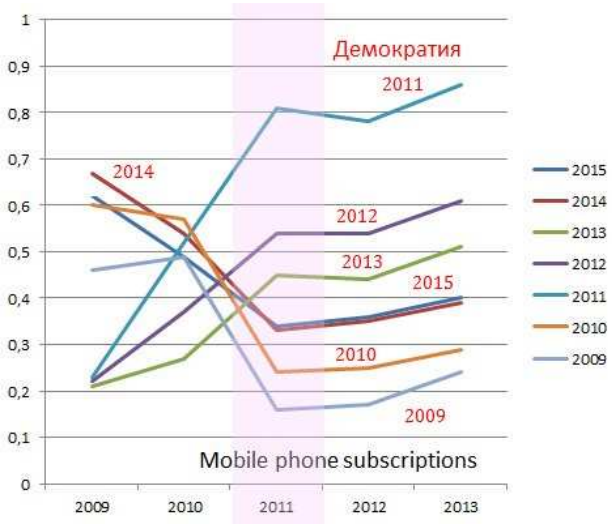
5-E



5-F



5-G



5-H

Рис. 5. Корреляционный анализ (значения r по Пирсону) для характеристик Big Data «терроризм» (A, B), «террорист» (C, D), «насилие» (E, F) и «демократия» (G, H) и статистического показателя «количество абонентов мобильной связи». Snail-структура.

При сравнении значений FSI в позитивной и негативной по имиджу группах было выявлено отличие: в позитивной группе среднее арифметическое FSI по суммарным очкам составило 81,3±4,9, в негативной группе среднее арифметическое FSI по суммарным очкам составило 94,0±14,5 (различия недостоверны). Все же, можно сказать, что страны, определенные в негативную по имиджу группу согласно проводимому анализу характеристик Big Data, имеют более высокий риск нестабильности, чем страны позитивной по имиджу группы. При сравнении стран по группам, разделенных по динамике экономических характеристик, отличий по FSI не выявлено.

5. Заключение

1. Массивы ключевых слов, относящиеся к категории Big Data и создающиеся хаотично глобальной Интернет-аудиторией, отражают в информационной среде Интернета реальные процессы, происходящие в глобальном социуме. Массивы ключевых слов можно использовать для прогностической оценки состояния государств. Датафицированные массивы ключевых слов коррелируют с классической статистической информацией, касающейся экономических и социальных сфер деятельности государств.

2. Выявлена пространственно-временная структура как отражение в Интернете политического явления «Арабская Весна», во взаимозависимости с распространением и использованием населением мобильных телефонов. Пространственно-временная структура схожа для всех 4-х анализируемых характеристик и условно по внешнему виду напоминает образ «улитки», что

позволило обозначить ее как Snail-структуру. Данные исследования позволяют сделать вывод, что такой глобальный политический процесс, как Арабская Весна, вмещающая в себя как демократические протестные движения, так и насилие и терроризм, была индуцирована насыщением региона MENA мобильными телефонами, а не влиянием соцсетей, как зачастую принято считать. С другой стороны, Интернет и соцсети в настоящее время переходят в сектор мобильных приложений. Скорость обмена информацией в Интернете между людьми существенно увеличивается, что можно считать фактором, усиливающим влияние Интернета и соцсетей на политические процессы в социуме.

3. Датафицированные массивы ключевых слов Интернета могут быть не только индикаторами социальных процессов, позволяя оценивать их интенсивность, но и при сопоставлении с классическими статистическими данными способствуют выявлению катализаторов социальных процессов. Данное исследование позволило определить в качестве катализатора персональную мобильную связь.

4. На основе доминирования той или иной характеристики в массивах ключевых слов каждая из анализируемых стран была определена в позитивную («развитие» /development/, «демократия» /democracy/) или негативную («террорист» /terrorist/, «наркотики» /narcotic/) группы по имиджу. Понятия «позитивная» и «негативная» группы надо воспринимать, как «зеркальное» отражение имиджа страны в глобальном информационном поле. Страны, определенные в негативную по имиджу группу согласно проводимому анализу характеристик Big Data, имеют более высокий риск нестабильности, чем страны позитивной по имиджу группы.

5. Обнаружены признаки посткризисной смены технологического уклада, которая обозначена как смена лидирующего тренда «Нефть» на тренд «ИТ». Смена технологического уклада характеризуется переходом на информационные технологии и солнечную энергетику, которые ассоциируются с развитием и позитивным имиджем государств.

6. Текстовая аналитика Big Data, как приоритет в области HPDA, требует углубления подходов с включением лингвистики, психологии, социологии, глобалистики, политологии. Простой морфологический подход (когда Интернет рассматривается как неструктурированная среда, постоянно заполняемая печатными словами) является весьма результативным. Необходимо и усложнение исследования, с установлением графов и кластеров.

Список литературы

[1] Science Landscape Seminar Series: Representative Big Data, Supercomputing and E-Infrastructure. June 2015. Council for Science and Technology. UK. [Электронный ресурс]. URL: https://www.gov.uk/government/publications/science-landscape-seminar-big-data-e-infrastructure-and-supercomputing (дата обращения: 12.07.2015).

[2] Next generation Computing and Big Data analytics. April, 2013. Report of House of Representatives, Subcommittee on Research & Subcommittee Technology Committee on Science, Space, and Technology, Washington, D.C. USA. 105 p.

[3] Доклад-презентация: Интеллектуальная метапоисковая система Sirius// Исследовательский центр искусственного интеллекта, Институт программных систем РАН. Переславль-Залесский, Россия. 2006 г. [Электронный ресурс]. URL: http://skif.pereslavl.ru/psi-info/airec/index.ru.html (дата обращения: 12.07.2015).

[4] Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data/ EMC Education Services. David Dietrich, Barry Heller, Beibei Yang. Published by John Wiley & Sons, Inc. USA. 2015. 435 p.

[5] Мировая динамика: Пер. с англ./ Д. Форрестер. М.: Изд-во АСТ; СПб.: Terra Fantastica, 2003. 379 с.

[6] Index of Economic Freedom. The Heritage Foundation. [Электронный ресурс]. URL: http://www.heritage.org/index/ (дата обращения: 28.04.2015).

[7] Networked Readiness Index. World Economic Forum. Global Information Technology Report 2015. [Электронный ресурс]. URL: http://reports.weforum.org/global-information-technology-report-2015/ (дата обращения: 28.04.2015).

[8] Fragile States Index. Fund for Peace. [Электронный ресурс]. URL: http://global.fundforpeace.org/ (дата обращения: 28.04.2015).

[9] Paul Comeau. Data-driven enquiry for Evidence-based Decision Making// DRDC CORA. Presentation to the 6th NATO Operational Analysis Conference, NC3A, The Hague, NL. June 2012.

[10] Ильин И.В., Каверин М.А. Вопросы преобразования международных организаций в институты глобального управления// Век глобализации. 2014. № 2. С. 32-38.

[11] Cukier K.N., Mayer-Schoenberger V. The Rise of Big Data// Journal Foreign Affairs. 2013. Vol. 92. No. 3. [Электронный ресурс]. URL: https://www.foreignaffairs.com/videos/2013-04-22/foreign-affairs-focus-kenneth-cukier-big-data (дата обращения: 28.04.2015).

[12] Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим/ Пер. с англ. Гайдюк И. М.: Манн, Иванов и Фербер, 2014. 240 с.

[13] Святогор Л., Гладун В. Семантический анализ текстов естественного языка: цели и средства. XVth International Conference «Knowledge-Dialogue-Solution», 2009. Ин-т кибернетики им. В.М. Глушкова НАН Украины. Kyiv, Ukraine. International Book Series «Information Science and Computing». Intelligent NL Processing. С. 9-18.

[14] Ильин И.В., Леонова О.Г., Розанов А.С. Теория и практика политической глобалистики. М.: Издательство Московского Университета, 2013. 296 с.

[15] Леонова О.Г. Прикладные аспекты политической глобалистики// Сборник материалов III Международного научного конгресса «Глобалистика-2013», посвященного 150-летию со дня рождения В.И. Вернадского. Москва, МГУ им. М.В. Ломоносова. 23-25 октября 2013 г. М.: МАКС Пресс, 2013. С. 263-265.

[16] Смородин Г.Н. Рынок трудовых ресурсов 2020// Академический форум корпорации EMC (EMC Academic Forum Russia & CIS). Сборник тезисов участников конференции. Москва, факультет ВМК МГУ им. М.В. Ломоносова. М.: МАКС Пресс, 2014 г. С. 3-5.

[17] Тоффлер Э. Третья волна: Пер. с англ./ Э. Тоффлер (Toffler A. The Third Wave, 1980). М.: Издательство АСТ, 2004. 781 с.

[18] Суперкомпьютер «Яндекс»// Журнал «Суперкомпьютеры», № 3(15), 2010. [Электронный ресурс]. URL: http://www.supercomputers.ru/ (дата обращения: 12.07.2015).

[19] Google cluster architecture. Report on Slide Share. 2011. [Электронный ресурс]. URL: http://www.slideshare.net/abhijeetdesai/google-cluster-architecture (дата обращения: 12.07.2015).

[20] Колесниченко О.Ю. XXI век: человеческое измерение и вызовы информационной глобализации. Монография. Saarbrücken, Germany, LAP LAMBERT Academic Publishing, 2015. 113 с.

Olga Kolesnichenko, Gennady Smorodin, Iliya Ilyin, Oleg Zhurenkov, Lev Mazelis, Dariya Yakovleva, Victor Dashonok. *Text's Big Data Analytics: perspective for supercomputers.*

Abstract. The article presents the results of the first phase of Big Data Analytics Multi-Center Study, which is organized by EMC Academic Alliance. Data Mining was carried out through use of multi-level hybrid scheme with access to non-classic supercomputers: Google and Yandex. Students who participated in the study gained practical skills during laboratory work, as well as Data Mining was included into the students' personal theses. The main question that was raised in this study concerned the assessment of what the Internet is in terms of numerous of different words that were written by people around the world as chaotic process. Is there a connection to real life of global society? It is shown that unstructured arrays of keywords in the Internet reflect the real processes in global society. Arrays of keywords can be used for prognostic assessment of the states.

Key Words and Phrases: Big Data, Data Mining, Text Analytics, non-classic supercomputers, Google, Yandex.

References

[1] Science Landscape Seminar Series: Representative Big Data, Supercomputing and E-Infrastructure. June 2015. Council for Science and Technology. UK. URL: https://www.gov.uk/government/publications/science-landscape-seminar-big-data-e-infrastructure-and-supercomputing.

[2] Next generation Computing and Big Data analytics. April, 2013. Report of House of Representatives, Subcommittee on Research & Subcommittee Technology Committee on Science, Space, and Technology, Washington, D.C. USA. 105 p.

[3] Report: Intelligent metasearch system Sirius // Research Center for Artificial Intelligence, Program Systems Institute RAS. Pereslavl, Russia. 2006. URL: http://skif.pereslavl.ru/psi-info/airec/index.ru.html.

[4] Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data/ EMC Education Services. David Dietrich, Barry Heller, Beibei Yang. Published by John Wiley & Sons, Inc. USA. 2015. 435 p.

[5] World Dynamics: Transcription from English /J. Forrester. Moscow Publishing house AST; St-Petersburg: Terra Fantastica, 2003. 379 p.

[6] Index of Economic Freedom. The Heritage Foundation. URL: http://www.heritage.org/index/.

[7] Networked Readiness Index. World Economic Forum. Global Information Technology Report 2015. URL: http://reports.weforum.org/global-information-technology-report-2015/.

[8] Fragile States Index. Fund for Peace. URL: http://global.fundforpeace.org/.

[9] Paul Comeau. Data-driven enquiry for Evidence-based Decision Making// DRDC CORA. Presentation to the 6th NATO Operational Analysis Conference, NC3A, The Hague, NL. June 2012.

[10] Ilyin I.V., Kaverin M.A. Questions of international organizations

transformation into the global governance institutions// Age of Globalization. 2014. № 2. pp 32-38.

[11] Cukier K.N., Mayer-Schoenberger V. The Rise of Big Data// Journal Foreign Affairs. 2013. Vol. 92. No. 3. URL: https://www.foreignaffairs.com/videos/2013-04-22/foreign-affairs-focus-kenneth-cukier-big-data.

[12] Mayer-Schoenberger V. Cukier K.N. Big Data: A Revolution That Will Transform How We Live, Work, and Think/ Transcription from English by Gaydyuk. Moscow: Mann, Ivanov and Ferber, 2014. 240 p.

[13] Svyatogor L., Gladun B. Semantic analysis of natural language texts: goals and methods. XVth International Conference «Knowledge-Dialogue-Solution», 2009. Glushkov Institute of Cybernetics. NASU. Kyev, Ukraine. International Book Series «Information Science and Computing». Intelligent NL Processing. 9-18 p.

[14] Ilyin I.V., Leonova O.G. Rozanov A.S. Theory and practice of political globalistics. Moscow: Publishing house of the Lomonosov Moscow State University, 2013. 296 p.

[15] Leonova O.G. Applied aspects of political globalistics// Proceedings of academic papers of III International Scientific Congress «Globalistics 2013», dedicated to the 150th anniversary of V.I. Vernadsky. Moscow, Lomonosov Moscow State University. October 2013. MAX Press, 2013. pp. 263-265.

[16] Smorodin G.N. Workforce 2020// EMC Academic Forum Russia & CIS. Proceedings of academic papers of Conference. Moscow, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University. Moscow. MAX Press, 2014 pp 3-5.

[17] Toffler A., The Third Wave. Transcription from English/ Alvin Toffler, The Third Wave, 1980. Moscow: AST Publishing, 2004. 781 p.

[18] The supercomputer «Yandex»// Journal «Supercomputers», № 3(15), 2010. URL: http://www.supercomputers.ru.

[19] Google cluster architecture. Report on Slide Share. 2011. URL: http://www.slideshare.net/abhijeetdesai/google-cluster-architecture.

[20] Kolesnichenko O.Yu. XXI Century: The Human Dimension and Challenges of Information Globalization. Monograph. Saarbrücken, Germany, LAP LAMBERT Academic Publishing, 2015. 113 p.

Об авторах – About authors

Ольга Юрьевна Колесниченко

Главный редактор информационного Бюллетеня «Анализ безопасности», PhD.
Область научных интересов – глобалистика, Большие данные.
E-mail: oykolesnichenko@list.ru

Olga Kolesnichenko
Editor in Chief of Security Analysis Bulletin, PhD, Moscow.
ORCID: 0000-0002-4523-6485

Геннадий Николаевич Смородин

Руководитель Академического Партнерства EMC в России и СНГ, PhD, MBA.
Область научных интересов – облачные технологии, Большие данные.
E-mail: gennady.smorodin@emc.com

Gennady Smorodin
Head of EMC Academic Alliance Russia & CIS, PhD, MBA, St. Petersburg.

Илья Вячеславович Ильин

Декан факультета глобальных процессов Московского Государственного Университета им. М.В. Ломоносова, профессор, доктор политических наук.
Область научных интересов – глобалистика, международные отношения, Большие данные.
E-mail: dekanat@fgp.msu.ru

Iliya Ilyin
Dean of the Faculty of global processes, Lomonosov Moscow State University, Professor, Doctor of Political Sciences, Moscow.

Олег Викторович Журенков

Доцент кафедры математики и прикладной информатики в экономике АНООВО «Алтайская академия экономики и права», PhD.
Область научных интересов – компьютерное моделирование, облачные технологии, Большие данные.
E-mail: zhur@pie-aael.ru

Oleg Zhurenkov
Department of Mathematics and engineering computer science in economics, Altai Academy of Economics and Law, Associate Professor, PhD, Barnaul.
ORCID: 0000-0003-4392-4134

Лев Соломонович Мазелис

Заведующий кафедрой математики и моделирования Владивостокского государственного университета экономики и сервиса, профессор, доктор экономических наук.
Область научных интересов – компьютерное моделирование, облачные технологии, Большие данные.
E-mail: lev.mazelis@vvsu.ru

Lev Mazelis
Head of the Department of Mathematics and Modeling, Vladivostok State University of Economics and Service, Professor, Doctor of Economic Sciences, Vladivostok.
ORCID: 0000-0001-7346-3960

Дарья Алексеевна Яковлева

Ассистент кафедры информационных технологий и систем Владивостокского государственного университета экономики и сервиса.
Область научных интересов – компьютерное моделирование, облачные технологии, Большие данные.
E-mail: Darya.Yakovleva330@yandex.ru

Dariya Yakovleva
Vladivostok State University of Economics and Service, Assistant of the Department of Information Technology and Systems, Vladivostok.
ORCID: 0000-0002-0139-4051

Виктор Леонидович Дашонок

Заместитель заведующего кафедрой «Информационные и вычислительные системы» ФГБОУ ВПО «Петербургский государственный университет путей сообщения Императора Александра I»; Академическое Партнерство EMC. Область научных интересов – облачные технологии, Большие данные.
E-mail: victor.dashonok@emc.com

Victor Dashonok
Deputy Head of Department of Information and Computing Systems, postgraduate, St. Petersburg State University of Railways of Emperor Alexander I, EMC Academic Alliance Russia & CIS, St. Petersburg
ORCID: 0000-0002-2803-251X

Слайд-презентация по докладу доступна по ссылке:

http://www.slideshare.net/OlgaKolesnichenko1/multienter-study-third-wave-big-data-2015

http://www.slideshare.net/OlgaKolesnichenko1/big-data-2015-52227865?related=1