ТЕОРИЯ, ИСТОРИЯ, МЕТОДОЛОГИЯ

DOI: 10.14515/monitoring.2015.5.02
УДК 303.442.3Академическое партнерство EMC

Правильные ссылки на статью:
Колесниченко О.Ю., Смородин Г.Н., Ильин И.В., Журенков О.В., Мазелис Л.С., Яковлева Д.А., Дашонок В.Л. «Третья волна»: многоцентровое исследование по аналитике Big Data Академического партнерства EMC в России и СНГ // Мониторинг общественного мнения: экономические и социальные перемены. 2015. № 5. С. 21-41
For citation:
Kolesnichenko O.Yu., Smorodin G.N., Ilyin I.V., Zhurenkov O.V., Mazelis L.S., Yakovleva D.A., Dashonok V.L. “The Third Wave”: Big Data Analytics multicenter study provided by EMC Academic Alliance in Russia & CIS // Monitoring of Public Opinion: Economic and Social Changes. 2015. № 5. Pp.21-41

О.Ю. КОЛЕСНИЧЕНКО, Г.Н. СМОРОДИН, И.В. ИЛЬИН, О.В. ЖУРЕНКОВ, Л.С. МАЗЕЛИС, Д.А. ЯКОВЛЕВА, В.Л. ДАШОНОК
«ТРЕТЬЯ ВОЛНА»: МНОГОЦЕНТРОВОЕ ИССЛЕДОВАНИЕ ПО АНАЛИТИКЕ BIG DATA АКАДЕМИЧЕСКОГО ПАРТНЕРСТВА EMC В РОССИИ И СНГ

«ТРЕТЬЯ ВОЛНА»: МНОГОЦЕНТРОВОЕ ИССЛЕДОВАНИЕ ПО АНАЛИТИКЕ BIG DATA АКАДЕМИЧЕСКОГО ПАРТНЕРСТВА EMC В РОССИИ И СНГ

КОЛЕСНИЧЕНКО Ольга Юрьевна – соискатель факультета глобальных процессов Московского государственного университета им. М.В. Ломоносова, главный редактор информационного бюллетеня «Анализ безопасности», PhD.
E-mail: oykolesnichenko@list.ru
ORCID: 0000-0002-4523-6485

СМОРОДИН Геннадий Николаевич – руководитель Академического партнерства EMC в России и СНГ, PhD, MBA. E-mail: gennady.smorodin@emc.com

ИЛЬИН Илья Вячеславович – декан факультета глобальных процессов Московского государственного университета им. М.В. Ломоносова, профессор, доктор политических наук.
E-mail: dekanat@fgp.msu.ru

ЖУРЕНКОВ Олег Викторович – доцент кафедры математики и прикладной информатики в экономике АНООВО «Алтайская академия экономики и права»,

“THE THIRD WAVE”: BIG DATA ANALYTICS MULTICENTER STUDY PROVIDED BY EMC ACADEMIC ALLIANCE IN RUSSIA AND THE CIS

KOLESNICHENKO Olga Yurievna – Degree-seeking student, Faculty of Global Processes, Lomonosov Moscow State University, Safety Analysis Information Bulletin Editor-in-Chief, PhD.
E-mail: oykolesnichenko@list.ru
ORCID: 0000-0002-4523-6485

SMORODIN Gennadii Nikolaevich – Head of EMC Academic Alliance in Russia and the CIS, PhD, MBA.
E-mail: gennady.smorodin@emc.com

ILIIN Iliya Vyacheslavovich – Dean of Faculty of Global Processes, Lomonosov Moscow State University, Professor, Doctor of Political Sciences.
E-mail: dekanat@fgp.msu.ru

ZHURENKOV Oleg Viktorovich – Associate Professor, Chair of Mathematics and Applied Informatics in Economy, Altai Academy of Economy and Law, PhD.

PhD. E-mail: zhur@pie-aael.ru
ORCID: 0000-0003-4392-4134

МАЗЕЛИС Лев Соломонович – заведующий кафедрой математики и моделирования Владивостокского государственного университета экономики и сервиса, профессор, доктор экономических наук.
E-mail: lev.mazelis@vvsu.ru
ORCID: 0000-0001-7346-3960

ЯКОВЛЕВА Дарья Алексеевна – магистрант Владивостокского государственного университета экономики и сервиса.
E-mail: malva-baksik@inbox.ru
ORCID: 0000-0002-0139-4051

ДАШОНОК Виктор Леонидович – заместитель заведующего кафедрой «Информационные и вычислительные системы» ФГБОУ ВПО «Петербургский государственный университет путей сообщения Императора Александра I»; Академическое партнерство EMC.
E-mail: victor.dashonok@emc.com
ORCID: 0000-0002-2803-251X

E-mail: zhur@pie-aael.ru
ORCID: 0000-0003-4392-4134

MAZELIS Lev Solomonovich – Head of Mathematics and Modelling Chair, Vladivostok State University of Economics and Service, Professor, Doctor of Economic Sciences.
E-mail: lev.mazelis@vvsu.ru
ORCID: 0000-0001-7346-3960

YAKOVLEVA Darya Alekseevna – Master`s Degree Student, Vladivostok State University of Economics and Service.
E-mail: malva-baksik@inbox.ru
ORCID: 0000-0002-0139-4051

DASHONOK Viktor Leonidovich – Deputy Head; Information and Computer Systems Chair, St. Petersburg State University of Communication; EMC Academic Alliance in Russia and the CIS, Russia. E-mail: victor.dashonok@emc.com
ORCID: 0000-0002-2803-251X

Аннотация. В статье представлены результаты первого этапа многоцентрового исследования по аналитике Больших данных, организованного по инициативе Академического партнерства EMC в России и СНГ. Показано, что неструктурированные массивы ключевых слов, относящиеся к категории Big Data, отражают в информационной среде Интернета реальные процессы, происходящие в глобальном социуме, и могут быть использованы для прогностической оценки состояния государств.
Например, датафицированная текстовая характеристика «мобильный телефон» в привязке к 2011 г., когда началась «арабская весна», оказалась связана с социально-демократическими процессами в глобальном обществе и не связана со статистическим увеличением числа пользователей мобильных телефонов в арабских странах. А статистический показатель «количество абонентов мобильной связи» коррелирует сразу с несколькими характеристиками Big Data:

Abstract. The article presents the results of the first phase of the Big Data Analytics Multicenter Study initiated by the EMC Academic Alliance in Russia and the CIS. The results show that unstructured arrays of keywords related to the Big Data reflect the actual global society processes in the Internet information environment. Arrays of special keywords can be used for prognostic assessment of the states in Big Data Analytics.
Dataficated text's words "mobile phone" related to 2011, when the Arab Spring started, turned out to be associated with social and democratic processes in the global society and not with a statistical increase in the number of mobile phone users in the Arab countries. A statistical measure of mobile phone subscribers correlates with several Big Data dataficated text's words such as "terrorist", "terrorism", "violence", and "democracy".
Contrary to the spread of mobile phones that contributed to the "Arab spring", no relationship between the number of the

«террорист» /terrorist/, «терроризм» /terrorism/, «насилие» /violation/, «демократия» /democracy/.

В противоположность распространению мобильных телефонов, оказавшему влияние на «арабскую весну», связи между такими статистическими показателями, как число пользователей Интернета и соцсетей, и перечисленными выше характеристиками Big Data, отражавшими общественные волнения в арабских странах, не обнаружено. Таким образом, полученные данные показывают, насколько важны средства мобильной связи в социальных процессах и насколько оправданны существующие подходы в борьбе с терроризмом, в которых основное внимание уделяется каналам мобильной связи террористов.

Выявлена пространственно-временная структура (поверхностная диаграмма корреляционного поля), отражающая политическое явление «арабская весна» в Интернете во взаимозависимости с распространением и использованием населением мобильных телефонов. Данная пространственно-временная структура схожа для всех 4-х анализируемых характеристик («террорист» /terrorist/, «терроризм» /terrorism/, «насилие» /violation/, «демократия» /democracy/). По внешнему виду она напоминает образ улитки, за что получила название «Snail-структура». В рамках заданных параметров исследования эта структура отражает сильную корреляционную связь между четырьмя вышеописанными характеристиками Big Data, привязанными к 2011 г., и распространением мобильных телефонов начиная с 2011 г. и далее. Можно сделать вывод о том, что насыщение региона персональной мобильной связью произошло к 2011 г., став одним из катализаторов массовых волнений арабского населения. Также обнаружена корреляционная связь с уровнем распространения мобильных телефонов в 2009 г., повторяющаяся для всех четырех датафицированных характеристик Big Data, в привязке к 2014–2015 гг. Объяснить такую ретроградную корреляционную связь пока затруднительно. Требуется продолжать исследование в данном направлении с опорой на полученные результаты.

Internet and social network users and the above-mentioned Big Data characteristics describing the Arab upheavals was revealed.

These findings strengthen the importance of mobile communications in social processes and the existing approaches to counteract terrorism where a special attention is paid to the terrorist mobile communication.

The spatiotemporal structure (surface diagram of correlation field) was defined to describe the Arab Spring as an Internet political phenomenon referring to the mobile phone penetration. This spatiotemporal structure is similar across four dataficated text's words ("terrorist", "terrorism", "violence", "democracy"). As it resembles a snail, it is called a Snail-structure.

The Snail-structure demonstrates a strong correlation between four dataficated text's words related to 2011 and mobile phones distribution in 2011 and later. It is concluded that the mobile phone penetration was so high in the MENA region that it became a catalyst for the Arab mass unrest. Another correlation was detected between dataficated text's words related to 2014-2015 and mobile phones' distribution in 2009. At present, it is difficult to explain this retrograde correlation. A further research based on the findings obtained is needed.

Ключевые слова: Большие данные, информационная глобализация, терроризм, «арабская весна», мобильный телефон, соцсети, датафикация, Snail-структура.

Keywords: Big Data, Data Mining, information globalization, terrorism, Arab Spring, mobile phone, social networks, datafication, Snail structure

В данном исследовании в центре внимания были неструктурированные текстовые Большие данные (Big Data), которые использовали 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 различным 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.

В сфере извлечения практической пользы из аналитики Больших данных есть успешный пример, который относится к оценке неструктурированных текстовых массивов, накапливающихся в Интернете – это проект Google Flu. Компания Google отслеживает динамику появления в Интернете определенных ключевых слов (в основном это запросы, вводимые людьми в окно поиска). Благодаря разностороннему анализу, специалисты Google определили 45 условий поисковых запросов, которые имеют высокий коэффициент корреляции с официальными эпидемиологическими данными по заболеваемости гриппом [Майер-Шенбергер, Кукьер, 2014; Cukier, Mayer-Schoenberger, 2013]. В результате в режиме реального времени удается определить, в каких регионах начинается эпидемия гриппа. Этот опыт аналитики Больших данных 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 можно извлечь важную информацию.

Неструктурированные текстовые массивы Больших данных в Интернете относятся к той категории данных, которые подходят для мониторинга ситуации. До появления Интернета человечество не имело опыта работы со словами в качестве данных. Тексты печатались на бумажных носителях, и не было возможности объединить то, что люди печатали или писали: не только статьи, но и телеграммы и письма. Теперь же тексты легко объединить в один массив при проведении Data Mining; 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 новостями, а также поисковыми запросами. *Интернет способствует тому, что речевая деятельность человека все больше переходит в письменную оцифрованную форму, т.е. Интернет датафицирует речевую функцию человека.* Если люди пишут, думают и говорят о чем-то больше или меньше, это должно иметь какие-то причины. **Слова** стали **данными**, анализируя которые можно получить информацию о текущей обстановке и сделать выводы о векторе развития ситуации.

Подчеркнем, что в исследовании не рассматривается так называемый Sentiment analysis (иначе Opinion mining) – подход, принципиально иной. Интернет рассматривается как

неструктурированная среда, постоянно заполняемая печатными словами. Интенсивность появления тех или иных слов имеет определенную динамику. При использовании поиска отражения в Интернете имиджа государств динамика накопления определенных слов может характеризовать обстановку в этих государствах.

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 О.Г. Леонова, профессор факультета глобальных процессов МГУ им. М.В. Ломоносова, среди деструктивных методов или глобальных политических технологий воздействия особое место занимают «мягкие» глобальные политические технологии и социокультурное воздействие (включая технологии управления массовым сознанием) [Леонова, 2013; Ильин, Леонова, Розанов, 2013]. Применение таких методов делает людей мишенью в глобальной перспективе, подрывая основу для устойчивого глобального развития. Кроме того, 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 факторов, катализирующих социальные процессы.

Методика

Исследование – уникальный проект по аналитике Больших данных (Big Data Analytics), впервые организованный в России по инициативе Академического партнерства EMC (EMC Academic Alliance). Одной из важных задач Академического партнерства EMC является формирование будущего рынка труда и создание прослойки высококвалифицированной рабочей силы, ориентированной на новые тенденции и технологии [Смородин, 2014].

В данной работе представлены результаты многоцентрового исследования, в котором приняли участие 3 вуза, входящих в Академическое партнерство EMC: Московский государственный университет им. М.В. Ломоносова, Алтайская академия экономики и права и Владивостокский государственный университет экономики и сервиса.

Исследование названо «Третья волна», по аналогии с термином, введенным американским политологом и философом Элвином Тоффлером [Toffler, 1980; Тоффлер, 2004]. «Третья волна» по Тоффлеру – это современная информационная эра, меняющая мир и способствующая процессам глобализации. Основное положение, сформулированное на этапе формирования плана исследования: *проводится анализ открытых интернет-ресурсов (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 по заданиям осуществляли 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 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. 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 по численности варьировали 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. Перечислим этапы получения результатов.

Case-study Big Data – описание цели исследования: оценка отражения в информационной среде имиджа государств и их экономических показателей посредством анализа текстов (ключевых слов из текстов).

Datafication Codebook – детальный перечень всех необходимых характеристик для извлечения данных. В исследовании это были текстовые ключевые слова (характеристики), а также название стран и годы, в которые публиковались тексты. Данные на английском и русском языках суммировались как один поток по каждой характеристике.

Математикам предоставляли файлы с перечнем слов, которые надо было фиксировать для подсчета их количества. Перечень ключевых слов был выбран исходя из опыта эмпирического анализа публикуемых текстов (как в СМИ, так и в соцсетях), в которых часто встречались подобные слова. Для анализа политической составляющей были определены следующие ключевые слова на русском и английском языках: «терроризм» /terrorism/, «террорист» /terrorist/, «оккупация» /occupation/, «наркотики» /narcotic/, «насилие» /violation/, «демократия» /democracy/, «развитие» /development/. Для анализа экономической составляющей были определены следующие ключевые слова на русском и английском языках: «капельное орошение» /drip irrigation/, «мобильный телефон» /mobile phone/, «солнечные батареи» /solar panel/, «атомная электростанция» /nuclear power plant/, «нефть» /oil/.

Было решено, что каждое ключевое слово в масштабах встречаемости, исчисляемой тысячами и миллионами (в привязке к названию той или иной страны и годам публикации текстов) будет отражать общую смысловую направленность публикуемых комментариев и текстов. Иначе говоря, например, упоминание слова «терроризм» /terrorism/ будет встречаться в статьях и комментариях, описывающих явления терроризма, в привязке к конкретной стране и году. Или слово «развитие» /development/ будет чаще встречаться в массивах текстов и комментариев в Интернете в привязке к конкретной стране и году, отражая позитивные процессы, происходящие в этой стране. Слово «мобильный телефон» /mobile phone/ будет встречаться в тех текстах и комментариях, которые описывают эту технологию мобильной связи и все, что с ней связано. Мобильная связь относится к технологическому тренду информационных технологий.

Исследование представляет собой первый шаг; в дальнейшем, опираясь на полученные данные, можно расширять перечень ключевых слов и даже попытаться составить шкалу релевантности тех или иных ключевых слов с точки зрения их неструктурированного накопления в Интернете и отражения политических и экономических процессов в странах. Конечной целью такой работы может быть отработка алгоритмов мониторинга на основе текстовых характеристик Big Data.

Сетка датафикации – детальное описание мэшапа (Mash-up, сопоставление разных блоков Больших данных). В данном исследовании сопоставление представляло следующую формулу: характеристика + (возможно, еще характеристика) + год публикации + название страны.

Первичная визуализация – по результатам мэшапа группой ИТ-специалистов составлялись графики и таблицы с датафицированными значениями.

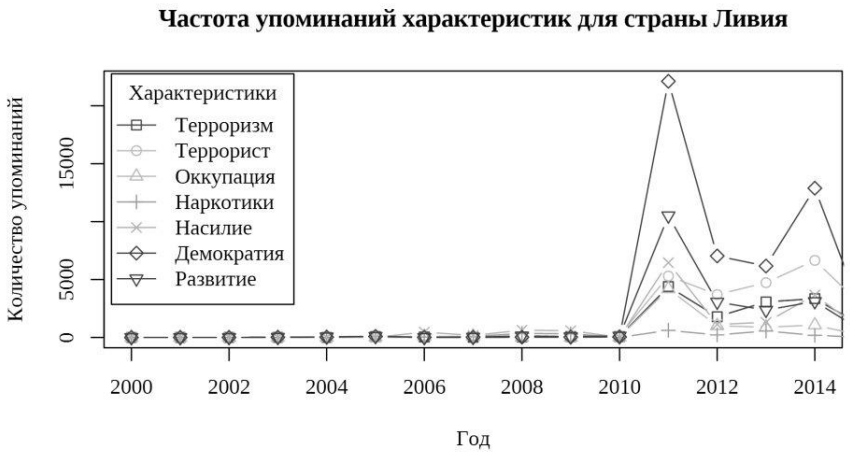
Вторичная визуализация – дополнительная графическая обработка результатов осуществлялась группой гуманитариев на основе предоставленных им табличных результатов, в целях выявления скрытых трендов и тенденций и описания результатов.

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.

Основные результаты и обсуждение

В данной статье будет представлена только часть полученных результатов. Из проанализированных ключевых слов Big Data наибольшую результативность показали следующие: «террорист» /terrorist/, «развитие» /development/, «демократия» /democracy/, «наркотики» /narcotic/, «мобильный телефон» /mobile phone/, «капельное орошение» /drip irrigation/, «солнечные батареи» /solar panel/, «нефть» /oil/. Данные характеристики могут быть заложены в алгоритмы, определяющие социальное и экономическое состояние государств, а также прогноз их развития. Несмотря на общую тенденцию постоянного увеличения объема слов в Интернете, погодовая динамика накопления ключевых слов имела разные тренды для различных стран, вошедших в анализ (рис. 1 а–е).

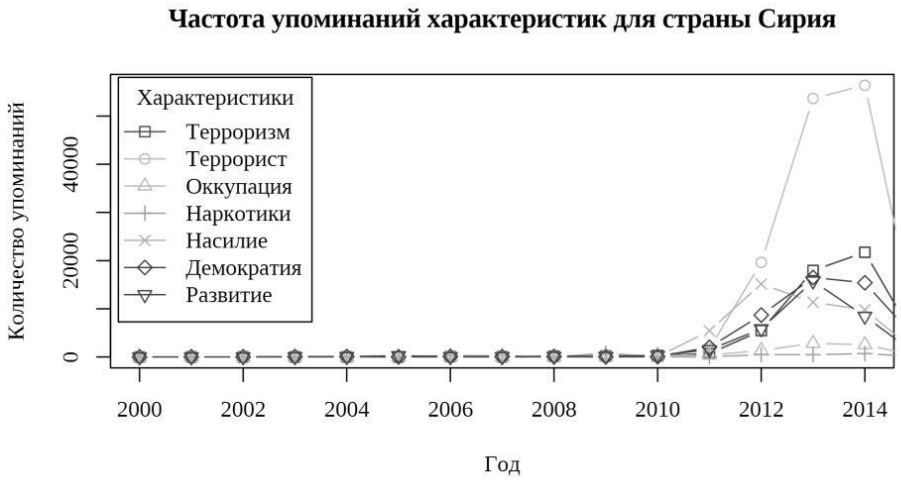
1а


Частота упоминаний характеристик для страны Ливия
Характеристики
Терроризм
Террорист
Оккупация
Наркотики
Насилие
Демократия
Развитие
Количество упоминаний
0
5000
15000
2000
2002
2004
2006
2008
2010
2012
2014
Год

1b


Частота упоминаний характеристик для страны Йемен
Характеристики
Терроризм
Террорист
Оккупация
Наркотики
Насилие
Демократия
Развитие
Количество упоминаний
0
1000
3000
5000
2000
2002
2004
2006
2008
2010
2012
2014
Год

1c


Частота упоминаний характеристик для страны Сирия
Характеристики
Терроризм
Террорист
Оккупация
Наркотики
Насилие
Демократия
Развитие
Количество упоминаний
0
20000
40000
2000
2002
2004
2006
2008
2010
2012
2014
Год

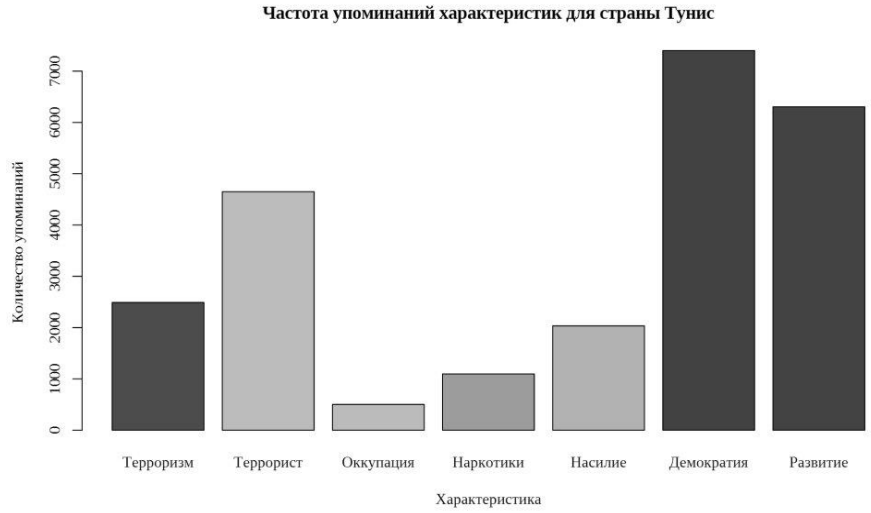
1d

1e

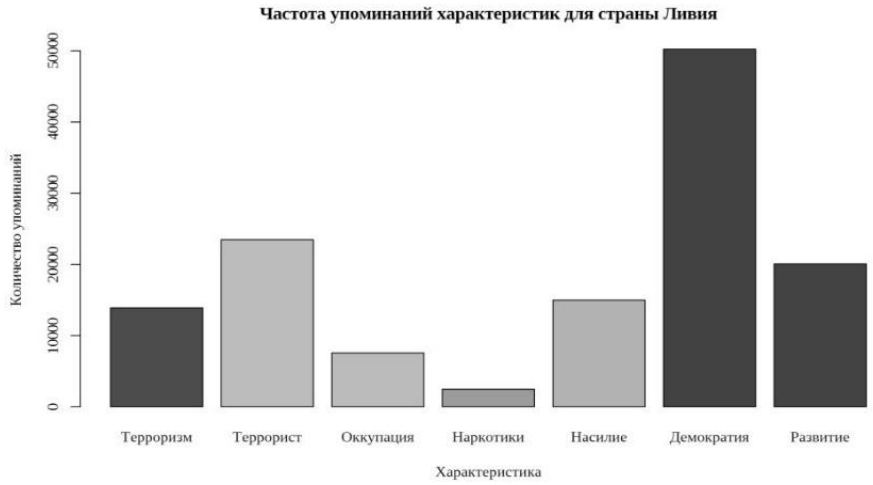
Рисунок 1 - Графики по результатам Data Mining в привязке к периоду 2000–2015 гг.: a – Тунис; b – Ливия; c – Йемен; d – Сирия; e – Ирак.

В анализ вошли 17 государств Евразийского континента и региона MENA (Middle East and North Africa – Средний Восток и Северная Африка): Азербайджан, Афганистан, Грузия, Израиль, Ирак, Иран, Йемен, Киргизия, Китай, Ливия, Пакистан, Палестина, Сирия, Тунис, Турция, Узбекистан, Украина. Перечень стран для исследования составляли на основе политической обстановки в них: были выбраны страны, в которых высок уровень терроризма, есть территориальные и религиозные конфликты. Так как рамки первого этапа исследования были ограничены возможностями вузов-участников по загрузке студентов на лабораторных работах, список стран не был исчерпывающим, что предстоит восполнить на следующих этапах исследования.

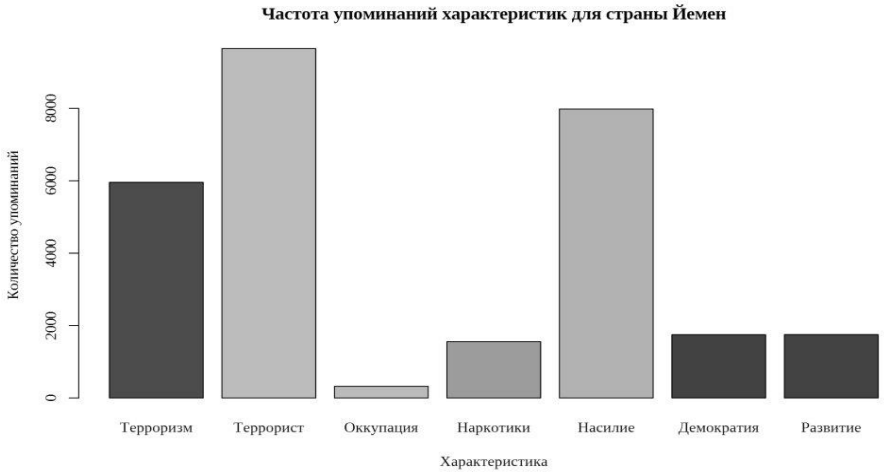
По результатам Data Mining, эти государства можно было разделить на две группы по имиджу: позитивная и негативная. Это распределение основывалось на суммировании характеристик за все годы, включенные в задание по извлечению данных (2000–2015 гг.), по каждому определенному ключевому слову (рис. 2 a–e).


Частота упоминаний характеристик для страны Тунис
Количество упоминаний
0
1000
2000
3000
4000
5000
6000
7000
Терроризм
Террорист
Оккупация
Наркотики
Насилие
Демократия
Развитие
Характеристика

2a

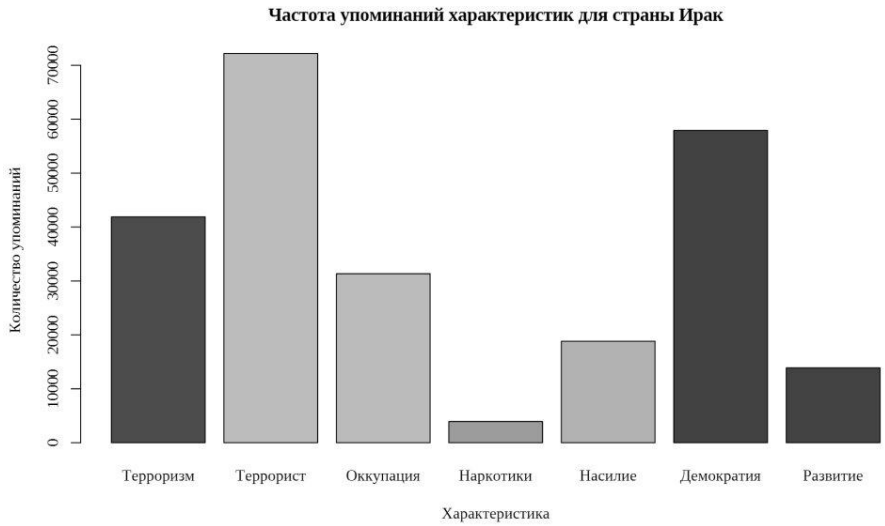

Частота упоминаний характеристик для страны Ливия
Количество упоминаний
0
10000
20000
30000
40000
50000
Терроризм
Террорист
Оккупация
Наркотики
Насилие
Демократия
Развитие
Характеристика

2b


Частота упоминаний характеристик для страны Йемен
Количество упоминаний
0
2000
4000
6000
8000
Терроризм
Террорист
Оккупация
Наркотики
Насилие
Демократия
Развитие
Характеристика

2c

2d



2e

Рисунок 2 - Графики по результатам Data Mining в привязке к 2000–2015 гг.: a – Тунис; b – Ливия; c – Йемен; d – Сирия; e – Ирак.

Из более подробно обсуждаемых в данной статье 5 государств, подвергшихся волне политических протестов, получившей общее название «арабская весна», к негативной по имиджу группе относились государства, у которых преобладала встречаемость характеристики «терроризм». К позитивной по имиджу группе относились государства, у которых преобладала встречаемость характеристики «демократия» (табл. 1 и 2).

Таблица 1 Хронология начала «арабской весны» в некоторых странах региона MENA

Тунис	17 декабря 2010 г.
Ливия	13 января 2011 г.
Йемен	18 января 2011 г.

Сирия	26 января 2011 г.
Ирак	10 февраля 2011 г.
Источник: https://ru.wikipedia.org.	

Таблица 2 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 г. к 2015 г.	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

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.

Так, Ливия (позитивная группа, «Демократия») перенесла войну, свержение и публичное убийство Муаммара Каддафи, но тема демократии стала лидирующим трендом в глобальном информационном поле 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). Именно с 18 декабря 2010 г., после самосожжения бедного продавца Мохаммеда Буазизи, начались массовые протесты сначала в Тунисе, а затем и в других арабских странах.

Графические данные Data Mining показали уникальный профиль для каждой страны отдельно. 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, а также отсутствие фильтрации являются важным условием исследования [Майер-Шенбергер, Кукьер, 2014; Cukier, Mayer-Schoenberger, 2013]. Для уточнения общей картины просматривались такие характеристики, как «насилие», «терроризм», «оккупация». На примере Ливии и Туниса видно, что все эти негативные признаки, указывающие на обсуждение отсутствия демократии, встречаются намного реже, чем характеристика «демократия» (рис. 1-a, 1-b, 2-a, 2-b).

Одним из компонентов анализа был FSI – Fragile States Index, ежегодно публикуемый вашингтонским исследовательским центром Fund for Peace [Fund for Peace, 2015]. При сравнении значений FSI в позитивной и негативной по имиджу группах стран было выявлено явное отличие: в позитивной группе среднее арифметическое FSI (по суммарным очкам) составило 81,3±4,9, в негативной группе среднее арифметическое FSI составило 94,0±14,5 (различия недостоверны). И 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.

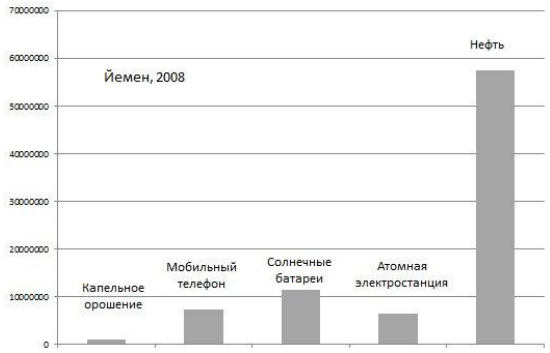
По результатам Data Mining массивов ключевых слов, указывающих на разные направления развития экономики (экономический блок ключевых слов анализировался в привязке к двум годам их публикации – кризисный 2008 и 2015 гг.), было выявлено, что в кризисный 2008 г. для всех стран (в анализ вошли 11 стран) по встречаемости лидировала характеристика «нефть» (группа «Нефть»). У некоторых стран к 2015 г. информационный фон изменился: лидирующей характеристикой стали ключевые слова «мобильный телефон», что позволило отнести эти страны в группу «ИТ (информационные технологии)» (рис. 3 a–j).

3a

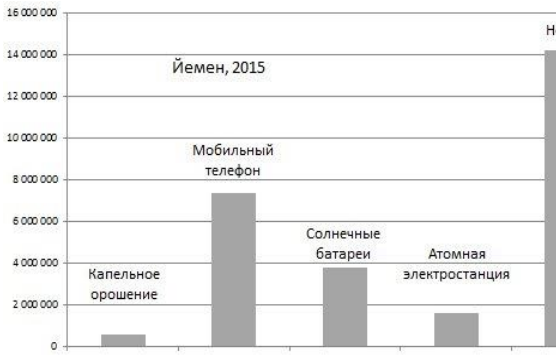
3b

3c

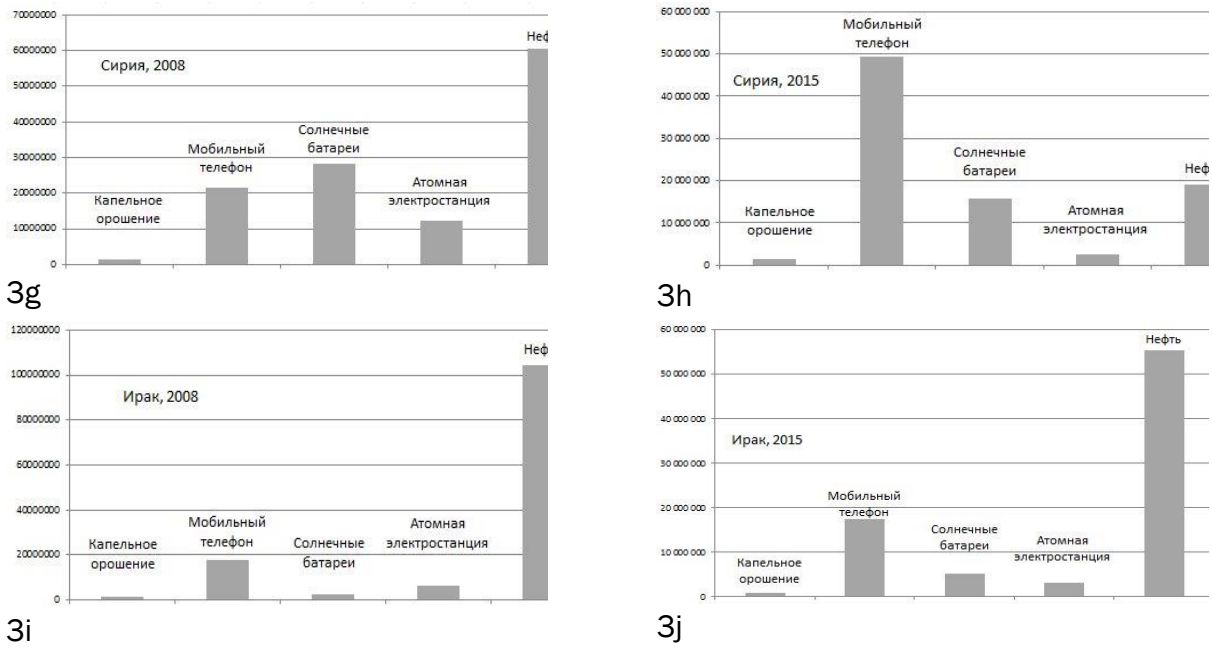
3d



3e



3f



3g 3h

3i 3j

Рисунок 3 - Графики по результатам Data Mining в привязке к 2008 и 2015 гг.: a, b – Тунис; c, d – Ливия; e, f – Йемен; g, h – Сирия; i, j – Ирак.

Сравнивая распределение 11 стран (Афганистан, Грузия, Ирак, Иран, Йемен, Ливия, Пакистан, Сирия, Тунис, Турция, Украина) по группам по имиджу (позитивная или негативная) и по экономическому тренду, удалось сделать интересные находки. Так, для стран, в которых отмечено изменение доминирующего тренда с 2008 по 2015 г. «Нефть»/«ИТ», среднее арифметическое по приросту 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, но для негативной по имиджу группы стран, составило 155±102%, что на 16% ниже среднего арифметического данного показателя для позитивной по имиджу группы стран, составившего 171±120% (различия недостоверны).

Данные иллюстрируют имеющуюся тенденцию, при которой в рамках темы устойчивого глобального развития и построения будущего, как правило, говорят о солнечной энергетике, в то время как тема нефти зачастую связана с негативными чертами современного мира: война, терроризм, борьба за ископаемые энергоресурсы [Колесниченко, 2015]. Это подтвердил и корреляционный анализ (коэффициент корреляции по Пирсону) между характеристиками Big Data «нефть» и «террорист», который показал для характеристики «нефть» в привязке к 2015 г. сильные положительные корреляционные связи с характеристикой «террорист», относящиеся к докризисному периоду (*r* изменяется от 0,7 до 0,8). Можно предположить, что в посткризисный период экономические подходы были пересмотрены, и *информационное поле Интернета отражает тот факт, что нефть стала ассоциироваться с негативными тенденциями и терроризмом*. Дополнительно можно отметить, что корреляционный анализ характеристик «нефть» и «развитие» не выявил ни одной сильной положительной корреляционной связи.

Данные, полученные в результате Data Mining, позволили заметить признаки

происходящей в настоящее время смены технологического уклада с постепенным переходом на информационные технологии и солнечную энергетику. При этом *энергетический переход позитивно влияет на имидж государств*. 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 возрастающих объемах.

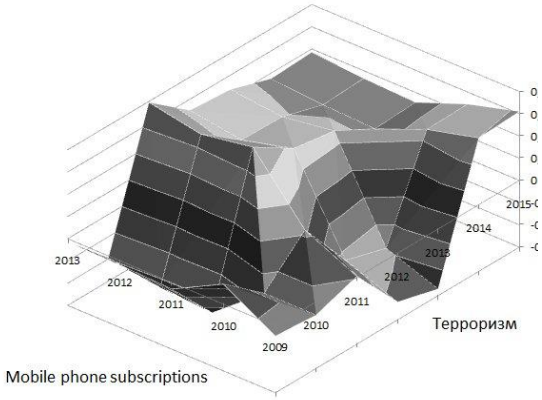
Было выявлено, что датафицированная характеристика «терроризм» имеет сильную отрицательную корреляционную связь (r= –0,75) со статистическим показателем ВВП на душу населения [Index of Economic Freedom… 2015]. Это свидетельствует в пользу того, что неструктурированные массивы слов, и в частности слово «терроризм», отражают 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. В то же время чем беднее государство, тем менее оно способно защитить себя от повторяющихся террористических актов. Еще одним доказательством того, что характеристика «терроризм» отражает реальные процессы, стало обнаружение достоверного отличия (F-тест, p<0,05) при сравнении значений Index of Economic Freedom (IEF) 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) – 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, т.е. 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» в информационном поле Интернета отражает важные социальные процессы в глобальном обществе. На это указывает обнаруженная сильная корреляционная связь с характеристикой «демократия»; с характеристиками «террорист», «терроризм», «насилие» были выявлены корреляционные связи средней силы. При этом не отмечено корреляции между частотой встречаемости характеристики Big Data «мобильный телефон» и статистическим показателем «количество абонентов мобильной связи» (Mobile phone subscriptions/100 pop) [Global Information Technology Report, 2015]. 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 исследованных странах.

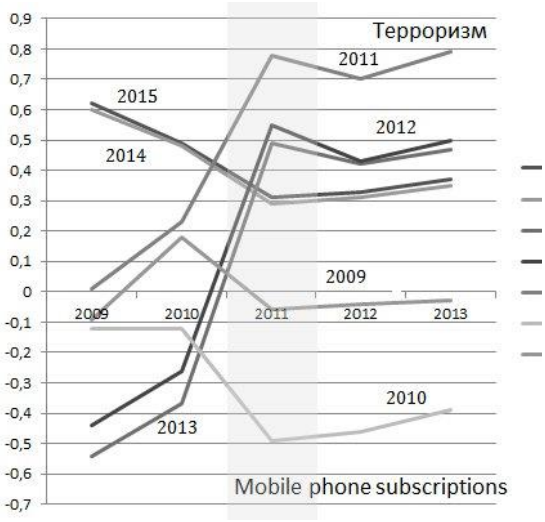
Были обнаружены корреляции статистического показателя «количество абонентов мобильной связи» (Mobile phone subscriptions/100 pop) сразу с несколькими датафицированными характеристиками Big Data: «террорист», «терроризм», «насилие», «демократия», при этом сильные положительные корреляции относились к 2011 г. Такое совпадение позволило предположить влияние «арабской весны» на корреляцию. Для подтверждения данного предположения был проведен более подробный корреляционный анализ, охватывающий перечисленные 4 характеристики Big Data и динамику статистического

показателя «количество абонентов мобильной связи» с 2009 по 2015 г. [The Global Information Technology Report, 2015]. И это предположение подтвердилось. 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 выпадение в тот или иной год статистических показателей по той или иной арабской стране. Для всех 11 стран, за единичным исключением в разные годы, имелись статистические данные по количеству абонентов мобильной связи.

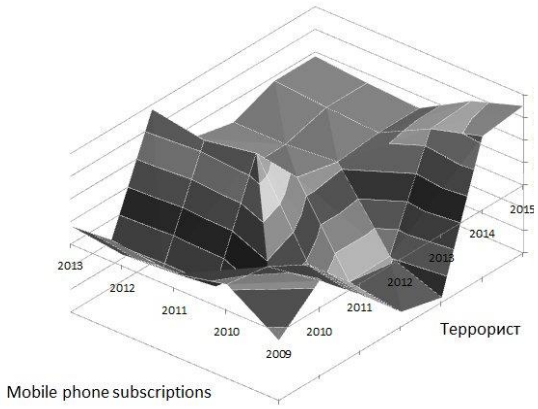
Рассмотрев поверхностные диаграммы, построенные как корреляционное поле, можно говорить о выявлении пространственно-временной структуры как отражения в Интернете политического явления «арабская весна» во взаимозависимости с распространением и использованием населением мобильных телефонов. По внешнему виду она напоминает образ «улитки» и может быть обозначена как Snail-структура (рис. 4 a–h).



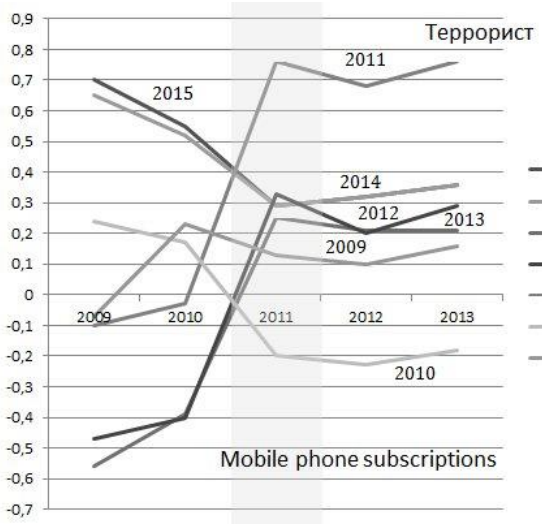
4a



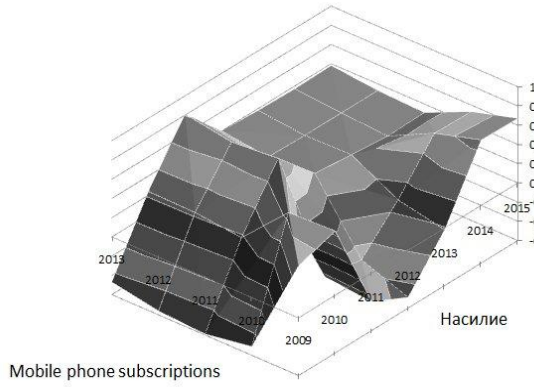
4b



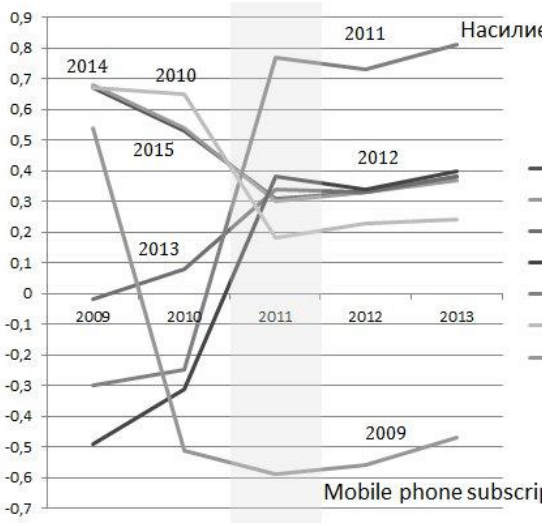
4c



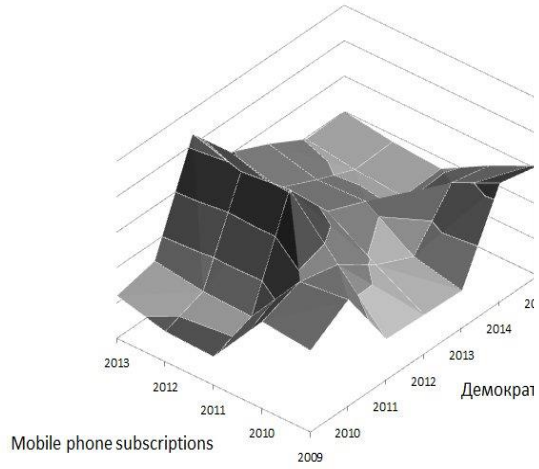
4d



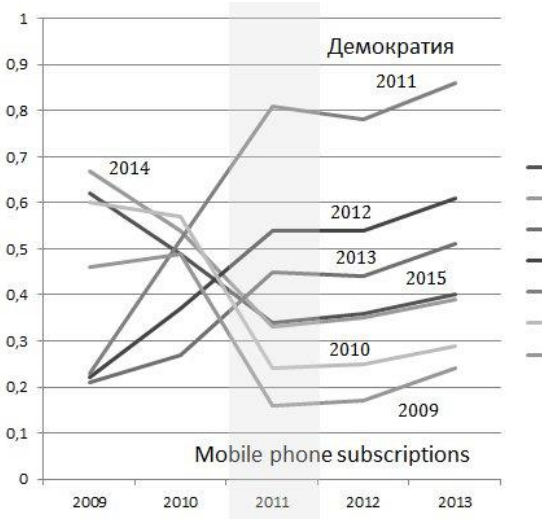
4e



4f



4g



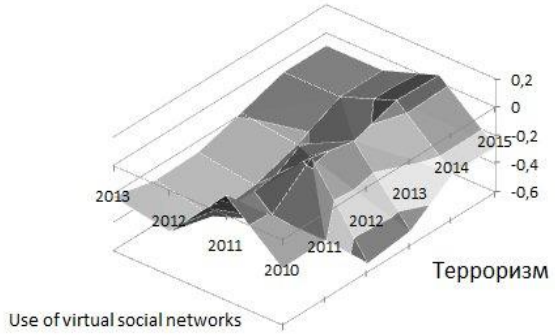
4h

Рисунок 4 - Корреляционный анализ (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».

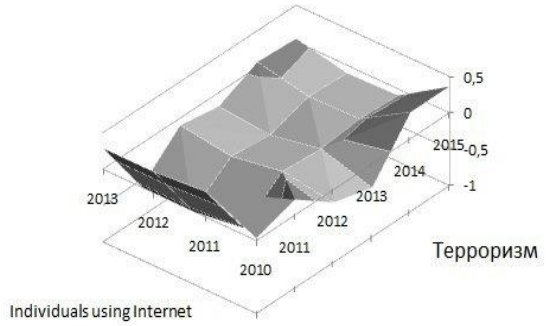
Сильные положительные корреляционные связи в привязке к 2011 г. обнаружены для характеристик «террорист» /terrorist/, «терроризм» /terrorism/, «насилие» /violation/, «демократия» /democracy/. Пространственно-временная Snail-структура схожа для всех четырех анализируемых характеристик. Находка исследования 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 четырьмя датафицированными характеристиками Big Data,

привязанными к 2011 г., и распространением мобильных телефонов начиная с 2011 г. и далее. Можно сделать вывод о том, что к 2011 г. произошло насыщение региона персональной мобильной связью, став одним из катализаторов массовых волнений арабского населения. Также обнаружена корреляционная связь, повторяющаяся для всех четырех датафицированных характеристик Big Data, привязанных к 2014–2015 гг., с уровнем распространения мобильных телефонов в 2009 г. Объяснить такую ретроградную корреляционную связь пока затруднительно. Требуется продолжать исследование в данном направлении с опорой на полученные результаты.

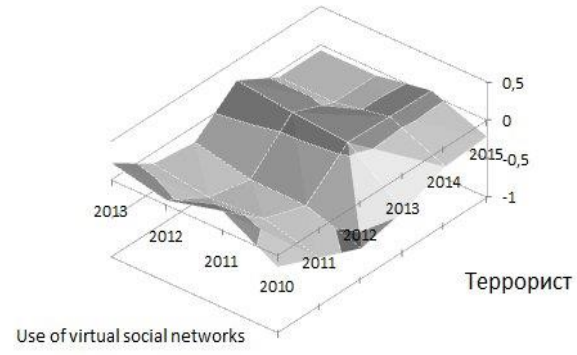
После получения корреляций, относящихся к статистическому показателю «количество абонентов мобильной связи» (Mobile phone subscriptions/100 pop), был проведен расширенный корреляционный анализ с такими статистическими показателями, как «число пользователей Интернета» (Individuals using Internet, %) и «вовлеченность населения в социальные интернет-сети» (Use of virtual social networks) [Global Information Technology Report, 2015]. Ни одной сильной корреляционной связи (т.е. достигнувшей значения 0,7 и выше) не выявлено, поэтому рассматривать эти поверхностные диаграммы с точки зрения их структуры не имеет смысла (рис. 5 ah).



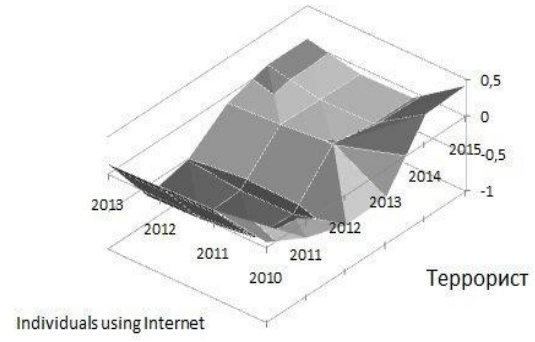
5a



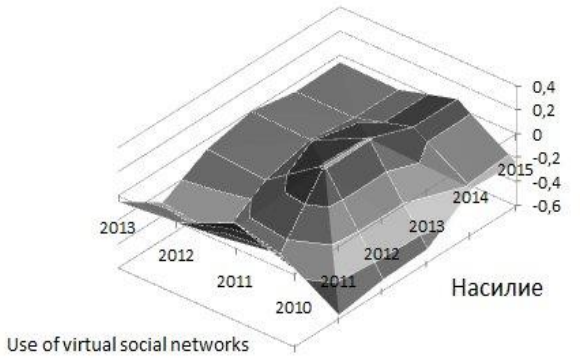
5b



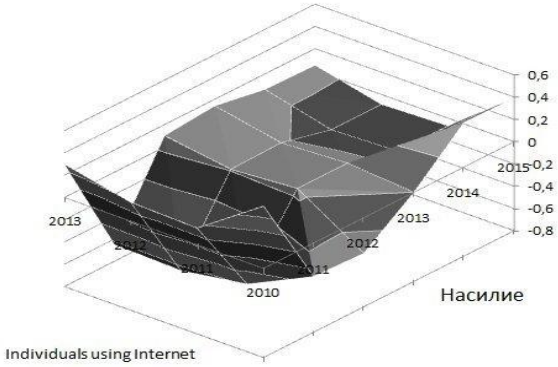
5c



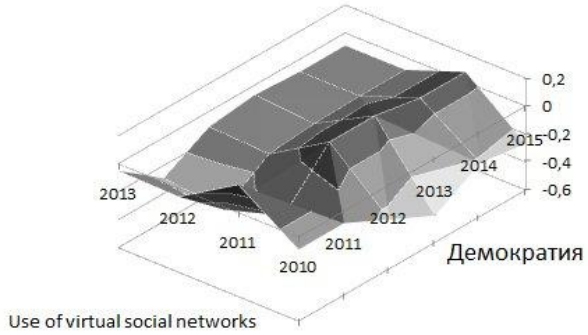
5d



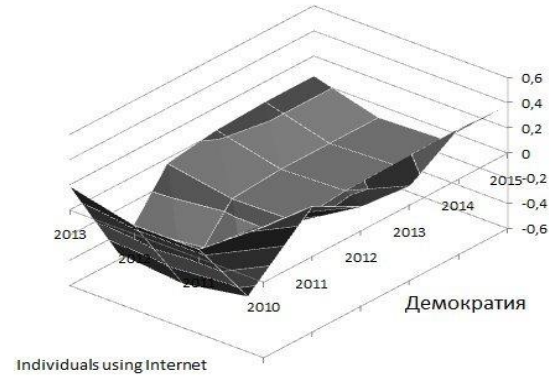
5e



5f



5g



5h

Рисунок 5 - Корреляционный анализ (по Пирсону) для характеристик Big Data «терроризм» (a, b), «террорист» (c, d), «насилие» (e, f) и «демократия» (g, h) и статистических показателей «вовлеченность населения в социальные интернет-сети» и «число пользователей Интернета».

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. Однако демонстрация этих данных в качестве диаграмм весьма интересна, так как любой опыт в аналитике Больших данных ценен с точки зрения разработки дальнейших исследований, в связи с чем научное сообщество призывает сообщать и о тех фактах, когда результаты не получены.

Выводы

Массивы ключевых слов, относящиеся к категории Big Data и хаотично создающиеся глобальной интернет-аудиторией, в информационной среде Интернета отражают 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.

Данные исследования позволяют сделать вывод, что такой глобальный политический процесс, как «арабская весна», вмещающий в себя как демократические протестные движения, так и насилие и терроризм, был индуцирован насыщением региона MENA мобильными телефонами, а не влиянием соцсетей, как зачастую принято считать. В то же время 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.

Датафицированные массивы ключевых слов Интернета могут быть не только индикаторами социальных процессов, позволяя оценивать их интенсивность, при сопоставлении с классическими статистическими данными они позволяют выявлять катализаторы социальных процессов. Данное исследование определило в качестве катализатора персональную мобильную связь, которая характеризуется тремя качествами: скоростью обмена информацией между людьми, возможностью обмениваться информацией в процессе любого действия, находясь в любом месте, и персональной адресацией сообщений, а также доверием к этой информации, получаемой от знакомого человека. Иными словами, мобильный телефон – это устройство, фактически прикрепленное к человеку и дополняющее его новыми возможностями для общения с другими людьми на больших расстояниях (в разы увеличивающее охват людей для личного доверительного общения). Стоит отметить, что современные технологические тренды направлены на создание не только более разнообразных интернет-приложений для мобильных телефонов, включая все соцсети и возможности передавать фотографии и видеоизображения с места событий, но и разработку более совершенных персональных мобильных (носимых на теле человека) устройств так называемой дополненной реальности (Augmented Reality), например, очков Google Glass, обладающих еще более мощным воздействием на человека. Таким образом, следует говорить о новом направлении в эргономике, которое совмещает в себе проблемы взаимодействия как отдельного человека и технического устройства, так и целого социума и технических устройств в аспекте индуцирования тех или иных социальных процессов.

Требуется продолжать исследовательскую работу по анализу текстовых массивов Big Data в направлениях, обозначенных в данном исследовании, с усложнением аналитических подходов и включением полного списка стран мира, используя кластерный анализ.

С точки зрения оценки протестного потенциала населения и прогноза в режиме реального времени локальных уличных столкновений рекомендовано проводить анализ аудио-, фото-, видео- и текстовых Больших данных, которые создаются при пользовании мобильными телефонами (с одновременным анализом территориальной локализации телефонов). Такая аналитика, в частности, проводится спецслужбами США и имеет много подводных камней в аспекте законности и нарушения прав пользователей мобильных телефонов.

Литература

1 Ильин И.В., Леонова О.Г., Розанов А.С. Теория и практика политической глобалистики. М.: Издательство Московского университета, 2013. 296 с.
2 Колесниченко О.Ю. XXI век: человеческое измерение и вызовы информационной глобализации. Монография. Saarbrücken, Germany, LAP LAMBERT Academic Publishing, 2015. 113 с.
3 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.
4 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.
5 Смородин Г.Н. Рынок трудовых ресурсов 2020 // Академический форум корпорации EMC (EMC Academic Forum Russia & CIS): Сб. тезисов участников конференции. Москва, факультет ВМК МГУ им. М.В. Ломоносова. М.: МАКС Пресс, 2014 г. С. 3–5.
6 Тоффлер Э. Третья волна. Пер. с англ. М.: Издательство АСТ, 2004. 781 с. (Toffler A. The Third Wave, 1980).
7 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).
8 Fragile States Index. Fund for Peace. [Электронный ресурс]. URL: http://global.fundforpeace.org/ (дата обращения: 28.04.2015).
9 Index of Economic Freedom. The Heritage Foundation. [Электронный ресурс]. URL: http://www.heritage.org/index/ (дата обращения: 28.04.2015).
10 Networked Readiness Index. World Economic Forum. Global Information Technology Report 2015. [Электронный ресурс]. URL: http://reports.weforum.org/global-information-technoogy-report-2015/ (дата обращения: 28.04.2015).