Большие Данные, Проблемы Общества и Цифровой Мир

Михаил Мягков, Консорциум Университетов

Университет Открытых Данных

28 апреля 2018 года

Университетский консорциум исследователей больших данных

Инфраструктура (данные, вычислительные мощности)

Методы (машинное обучение, нейронные сети, анализ естественного языка, сетевой анализ …)

Приложения (безопасность, медицина, образование …)

Структура Консорциума

Инфраструктура

В Томском государственном университете организован специальный серверный кластер для агрегирования гетерогенных данных.

Суперкомпьютер СКИФ Cyberia

- Пиковая производительность 107 Тфлопс
- >6000 вычислительных ядер
- Система хранения данных >400 Тб

Создан коллектив специалистов в области компьютерных технологий, которые понимают, каким образом нужно собирать, обрабатывать и хранить данные, и представителей других наук – социологов, политологов, филологов, психологов, биологов, генетиков.

Накоплены компетенции сотрудников по структурированию данных и выгрузке информации, отвечающей определенным критериям.

Данные

- Данные социальных сетей (Вконтакте, Одноклассники и др.), онлайн-СМИ, блогов, тематических и региональных форумов
- Данные по 650 000 контрактам госзакупок
- Дампы текстовой информации интернет-энциклопедии Wikipedia на 18 языках
- 170 000 идентифицированных страниц выпускников в интернет-энциклопедии Wikipedia из 326 университетов мира
- Профдиагностика 10000 школьников (1С «Профдиагностика»)
- Собственные данные университета

Приложения

Данные + Инфраструктура

Социально значимые

- Исследование проявлений благотворительности в онлайн-пространстве
- Оценка коррупционных рисков на основе данных государственных закупок
- Цифровое качество жизни
- Предсказание политических предпочтений пользователей социальных сетей

Образование

- Определение образовательных интересов и признаков одаренности у школьников
- Оценка влияния университетов на общество
- Анализ социально-психологического профиля личности

Коммерческие проекты

- Расчет показателей для Московского международного рейтинга вузов «Три миссии университетов»
- Маркетинговые исследования (“Триколор ТВ”, “Castrol”)
- Аналитические отчеты (МЧС России)

Безопасность

- Исследование экстремистских онлайн-сообществ в социальных медиа
- Анализ проявлений девиантного поведения среди школьников

Определение образовательных интересов и признаков одаренности у школьников

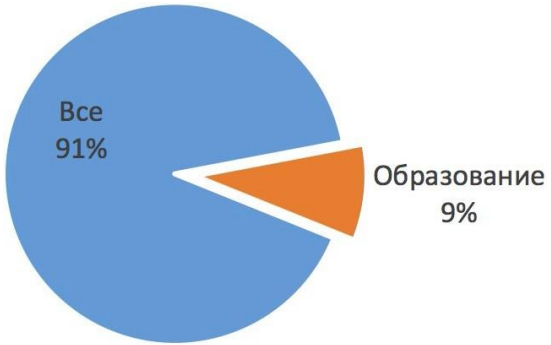
Выявление корреляции между образовательными интересами и когнитивными способностями школьника с одной стороны и его «электронным следом в социальной сети» с другой, на основе данных анализа профилей абитуриентов в социальной сети Вконтакте.

Данные:

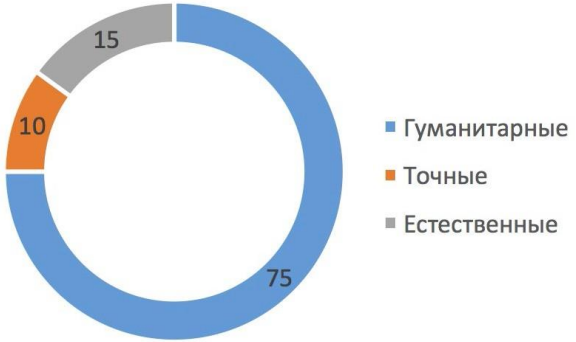
- Профили ВК 126 000 потенциальных абитуриентов СФО
- Контент анализ маркерных сообществ (150+)
- Классификация тематических сообществ (100 000+)
- Профдиагностика школьников 3000+

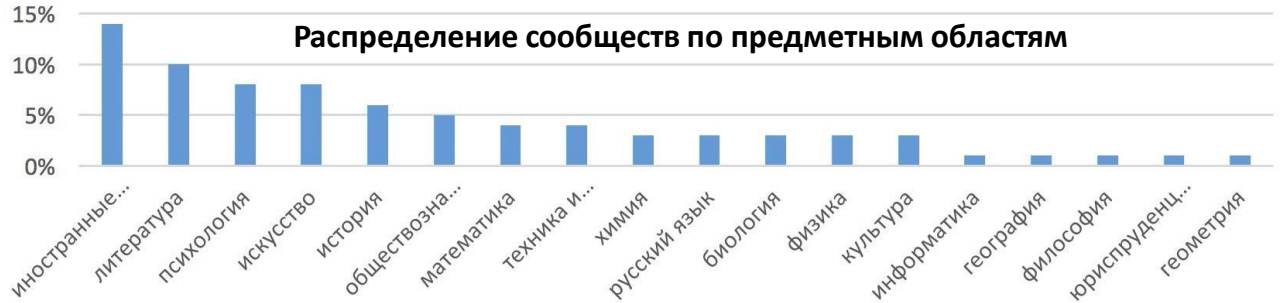
Образовательные интересы: анализ подписок

Профиль интересов



Профиль образовательных интересов





Образовательные интересы: предсказание профиля обучения

Выявлено 9000 гуманитариев

Приглашено 900 (точность 82%)

199 заявлений

56 зачислений

	Традиционный рекрутинг	Поиск и приглашения в ВК
Средний балл в аттестате	4,53	4,68
Средний балл ЕГЭ	212	224
Доля медалистов и отличников	27%	32%

Разработка показателей для оценки влияния университета на общество

Поиск выдающихся выпускников университета с использованием данных интернет-энциклопедии Wikipedia. Предложен показатель, учитывающий влияние университета на общество через выпускников, основанный на поиске выдающихся выпускников университета с использованием данных интернет-энциклопедии Wikipedia.

- Источник данных — данные страниц интернет-энциклопедии на национальном для университета и английском языках
- Поиск основан на применении лингвистических маркеров и технологий обработки естественного языка
- Формируется список выпускников университета, у которых имеется страница в Wikipedia
- Учитывается статистика посещаемости страницы выпускника для выделения выдающихся и наиболее влиятельных выпускников

Результаты

Идентифицировано 187 тысяч выпускников из 346 университетов. Данные переданы рейтинговому агентству Эксперт РА для расчета Московского международного рейтинга университетов.

ФИО выпускника	Университет	Количество просмотров страницы
Барак Обама	Columbia University	**16 673 705**
Владимир Путин	Saint Petersburg State University	**12 139 843**
Илон Маск	Queen’s University	**11 721 349**

Топ 3 выпускников по количеству просмотров за 2016 год

Результаты

В рейтинге 346 международных университетов из 39 стран мира.

Страна	Количество университетов/ количество выпускников*	Общее количество просмотров страниц выпускников за 2016 г.
Россия	50 унив. / 11 347 выпускников	**187.8 млн**
США	41 унив. / 52 460 выпускников	**988.49 млн**
Китай	21 унив. / 3 604 выпускника	**60.8 млн**
Великобритания	18 унив. / 13 533 выпускника	**350.8 млн**

* Включая исторических личностей, окончивших университет

Построение портрета выпускника

Анализ профилей, подписок, активности (лайки, репосты и комментарии) выпускников университетов с целью построение портрета выпускника:

Социальная ответственность	Социальные репосты Социальные группы и мероприятия
Иностранный язык	Информация из поля пользователя Подписки + активность там (лайки, репосты, коменты) Пишет на иностранном языке (объем текста – фиксировать)
Грамотность	Ошибки Использование мата Средняя длина слов
Коммуникабельность	Количество друзей Количество активных связей Анализ социального графа
Лидерство/инициативность	Администратор групп Создатель встреч, опросников Создание собственного контента
Креативность	Администратор групп с творческим направлением + посты с подписями Анализ текстов
Вовлеченность в жизнь страны	Репосты новостей (эмотивная оценка) Политические группы
Управление временем	Среднее количество часов в ВК в сутки
Саморазвитие/обучаемость	Подписки на образовательные, просветительные группы

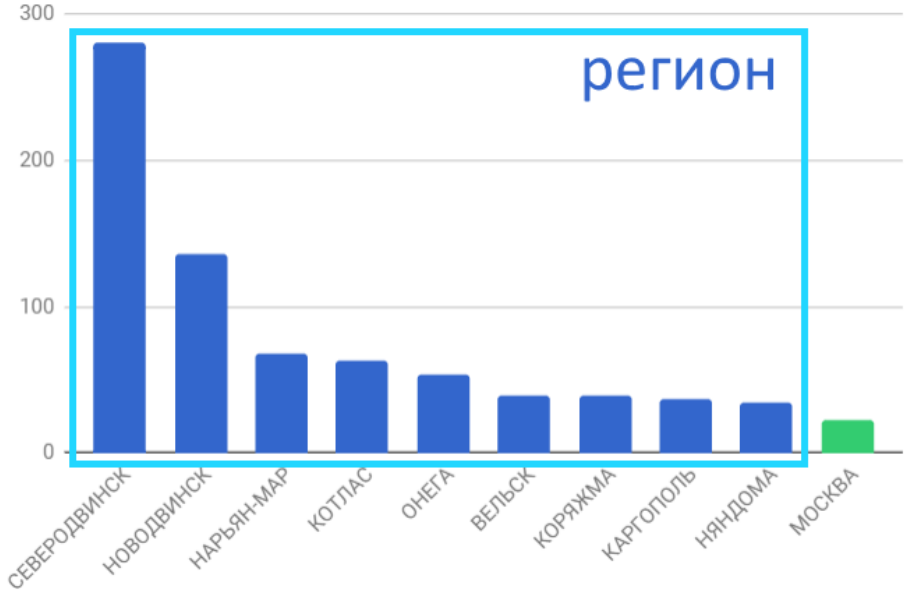
Выпускники САФУ

Идентифицировано 47 тысяч выпускников САФУ

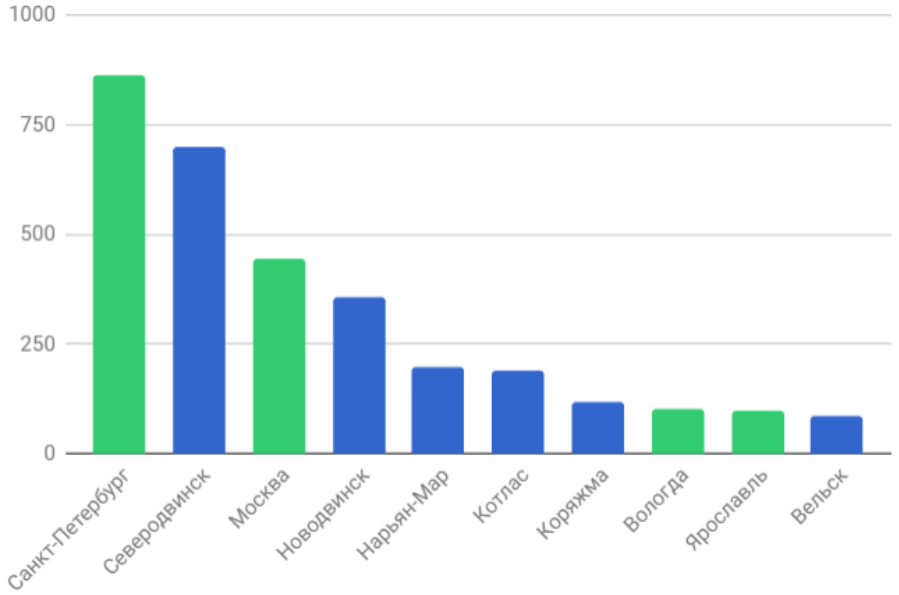
Выпускники САФУ – Миграция (топ 10 локаций)

область → университет → центр

Родной город



Сегодня



Архангельск занимает долю ~75% в обеих выборках Топ11

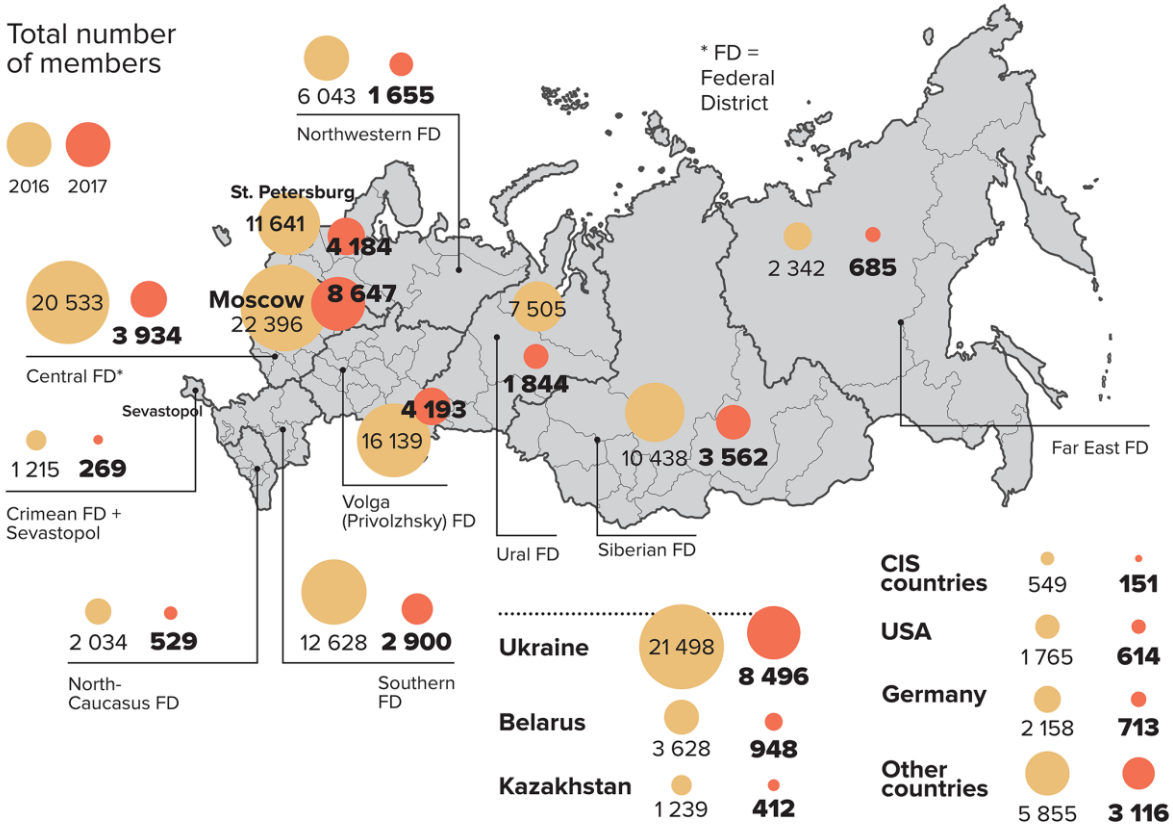
Исследование экстремизма в социальных медиа

Исследование характеристик, сетевой структуры и особенностей внутри- и мужгрупповых связей экстремистских сообществ в русскоязычном сегменте Интернет.

Идентифицировано 42 праворадикальных и 33 исламистских онлайн сообществ с общим количеством участников >860 000, 21 сообщество закрыто по решению судов РФ за время проведения исследования.

- > 2 млн профилей пользователей
- > 1.7млн лайков, 350 тысяч репостов и 1 млн комментариев
- 17 млн связей (совместная дружба) для оценки внутригрупповых связей
- классификация 417 тысяч групп для оценки перекрестных репостов контента

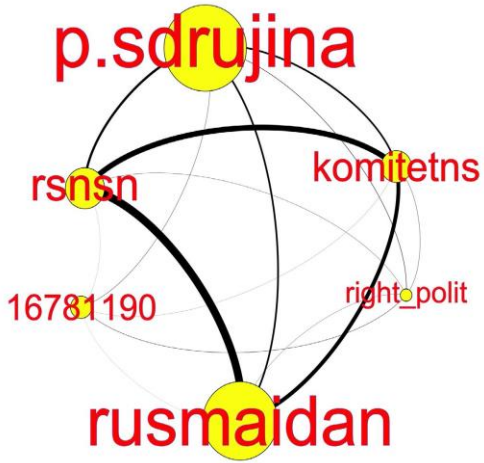
Географическое распределение участников идентифицированных праворадикальных онлайн сообществ в 2016-2017 г.



Динамика внутренней сетевой структуры

2016
Топология Star-structure

2017
Децентрализованная топология

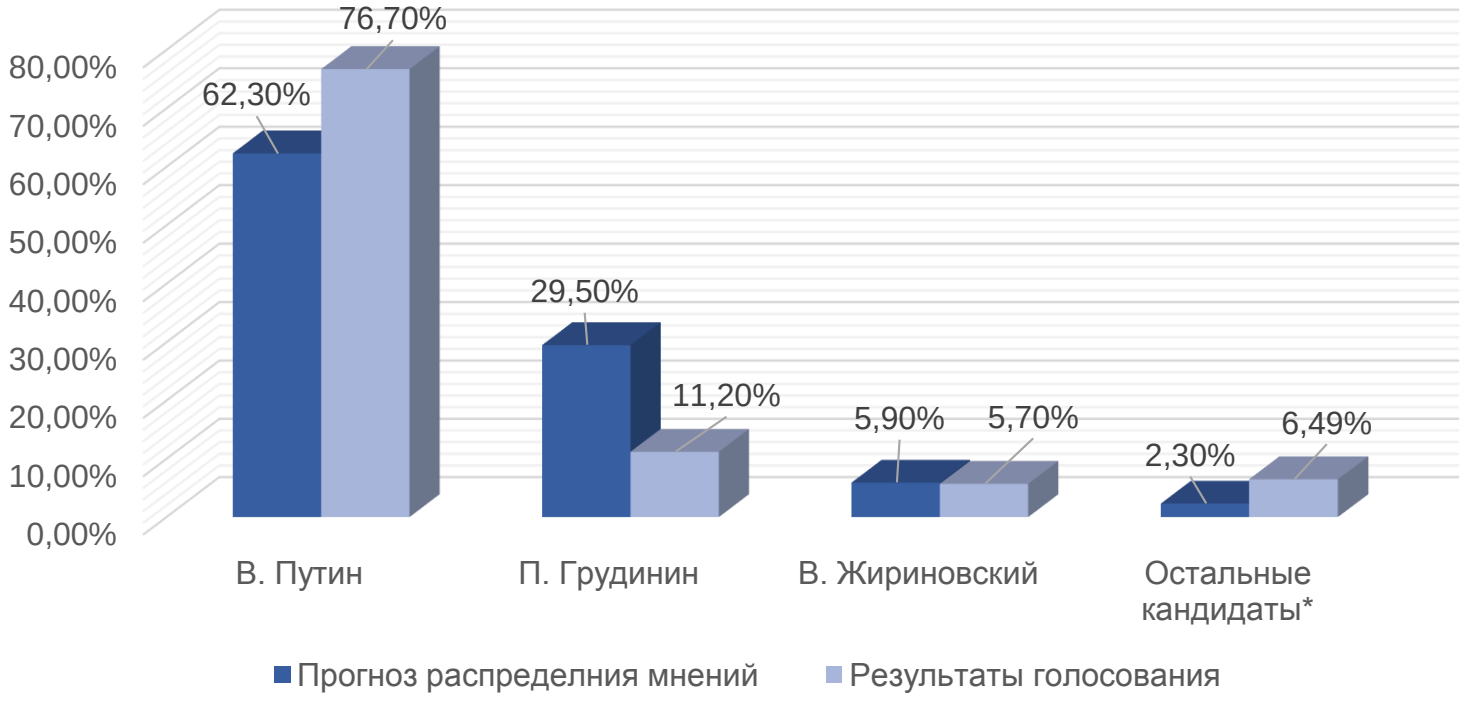


Прогнозирование политических предпочтений пользователей социальных медиа

Политические взгляды пользователей проецируются на их аккаунты в социальных сетях, пользователь будет подписываться на источники информации исходя из своих политических предпочтений.

С использованием трехслойной искусственной нейронной сети (ANN), проведен анализ и классификация пользователей в соответствии с их политическими предпочтениями. В качестве обучающей выборки выступали данные открытого опроса, проведенного среди 34000 пользователей Вконтакте.

Оценка политических взглядов 23 млн. пользователей Вконтакте



Оценка нерегулярностей в данных госзакупок и связанных с ними коррупционных рисков

Разработка Взвешенного индекса коррупционного рисков (ВИКР) на основе анализа исторических данных о проведенных конкурсах, введения индикаторов коррупционного риска и использовании логистической регрессии. Методология расчета индекса основана на идее, описанной в работе М. Фазекаса [1], но модернизирована для применения в России.

Данные

- выгрузка Единой информационной системы в сфере закупок РФ
- 652 882 контракта на общую сумму 11.3 триллиона рублей
- за период 2011–2017 гг.
- сумма каждого контракта >100 000 рублей

1. M. Fazekas, J. Chvalkovska, J. Skuhrovec, I.J. Tóth, L.P. King. 2013. Anatomy of grand corruption: A composite corruption risk index based on objective data.

Новые проекты

Идентификация социально-психологического профиля личности абитуриента и студента, прогнозирующего профессиональную успешность

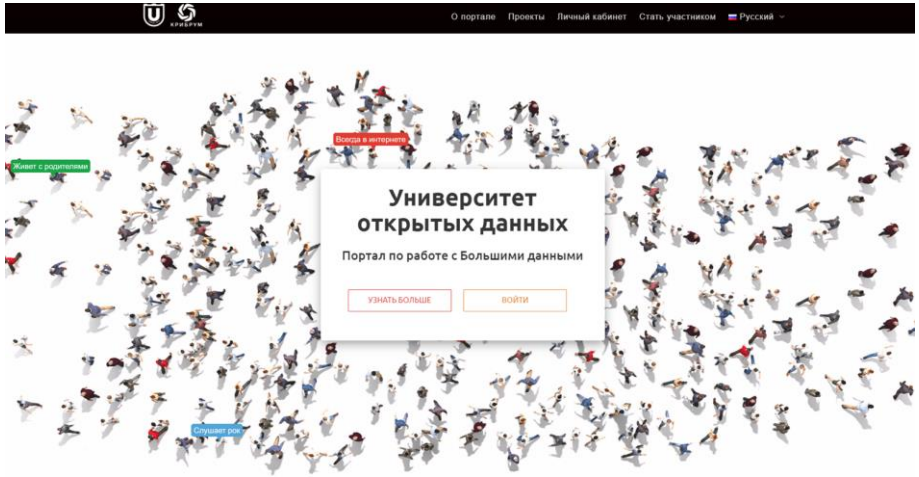
Определение социально-психологического профиля личности, отражающего её возможности обучения и профессиональной самореализации. Для определения психотипа используется методы психосемантического графа и моделирования коммуникативных миров на основе данных социальных сетей.

Идентификация благотворителей и исследование проявлений благотворительности в онлайн-пространстве

Выявление объектов сферы благотворительности и заинтересованных стейкхолдеров в сети. Разработка правил работы с разными возрастными группами (потенциальными благотворителями), выявление особенностей возрастных групп, выявление самой заинтересованной группы с точки зрения поведения благотворителя.

Университет открытых данных

Портал по работе с большими данными data.tsu.ru



Возможности Портала

- Доступ к данным (СМИ, социальные сети, блоги, цифровые следы пользователей, местоположение и т.д.) и обмен ими
- Методы и технологии обработки и анализа данных
- Совместные междисциплинарные исследования с объединением научных коллективов и компетенций
- Сотрудничество с коммерческими компаниями
- Доступ к вычислительным мощностям

Михаил Мягков

руководитель лаборатории наук о больших данных и проблемах общества Томского государственного университета, профессор университета штата Орегон (США)

myagkov@uoregon.edu
+7 916 825 9969