ИСП РАН

Анализ социальных медиа: задачи и технологические вызовы

Арутюн Ишханович Аветисян

Денис Юрьевич Турдаков

Математические методы распознавания образов
26 ноября 2019

ИСП РАН

Цифровая экономика: приложения

Непрерывный доступ в сеть интернет/интранет

Киберфизические возможности

Большая вычислительная сложность

Умные устройства (дом, офис, завод)

Платформы «интернета вещей», искусственного интеллекта, …

Платформы хранения и обработки «больших» данных

Облачные платформы

Аппаратура

Проблемы современного системного ПО:

1.Эскалация размеров (Astra Linux – более 150 миллионов строк кода).
2.Сложность среды разработки и сборки.
3.Отсутствие изолированных систем.

Необходимые качества системного ПО:

Эффективность

Продуктивность

Безопасность

Социальные медиа

Источники

- Социальные сети
- Веб-форумы
- Блоги
- Пользовательские комментарии
- Социальная журналистика

Задачи

- Маркетинговые, социологические исследования
- Конкурентная разведка
- Ситуационный мониторинг
- Управление репутацией публичных персон и организаций
- Политологические исследования
- Рекомендательные системы
- Персонализированная реклама
- Правоохранительные мероприятия

Вызовы при анализе социальных медиа

Большие данные

- Большой объем связанной между собой информации
- Поток только текстовых сообщений – сотни ГБ в день
- Некоторые закономерности можно заметить, только на выборках сравнимых по размеру с генеральной совокупностью (все информационное пространство)
- Разнородная информация, требующая интеграции

Специфичный язык

- Сленг, сокращения, отсылка к внешнему контексту
- Высокая скорость изменения
- Сочетание текстового, графического и аудио контента

Недостоверность информации

- Боты
- Анонимность, неверно указанная и неполная информация в профилях
- Дезинформация

Исследования 2012-2014: пакетная обработка

Вызовы

- Обнаружение сообществ в социальных сетях
  - 1 млрд. вершин,
  - 100 миллиардов ребер,
  - за 24 часа
- Проведение экспериментального тестирования и доказательство масштабируемости

Решения

- Разработан итеративный алгоритм на основе распространения меток
- Набирающий популярность стек технологий Apache Hadoop не подходил для итеративных алгоритмов (реализация модели MapReduce предполагала чтение/запись в распределенную файловую систему HDFS на каждой итерации)
- Найдена альтернатива: (Apache) Spark (UC Berkeley) с помощью которой решена задача
- Разработаны модель и алгоритм генерации графов большого размера со свойствами социальных сетей

1. K. Chykhradze, A. Korshunov, N. Buzun, R. Pastukhov, N. Kuzyurin, D. Turdakov, H. Kim. Distributed generation of billion-node social graphs with overlapping community structure / // Complex Networks V. — Springer, 2014.
2. N. Buzun, A. Korshunov, V. Avanesov, I. Filonenko, I. Kozlov, D. Turdakov, H. Kim. EgoLP: Fast and distributed community detection in billion-node social networks / // Data Mining Workshop (ICDMW) — IEEE. 2014.

Исследования 2015-2017: потоковая обработка

Вызовы

Обнаружение информационных кампаний на ранних этапах

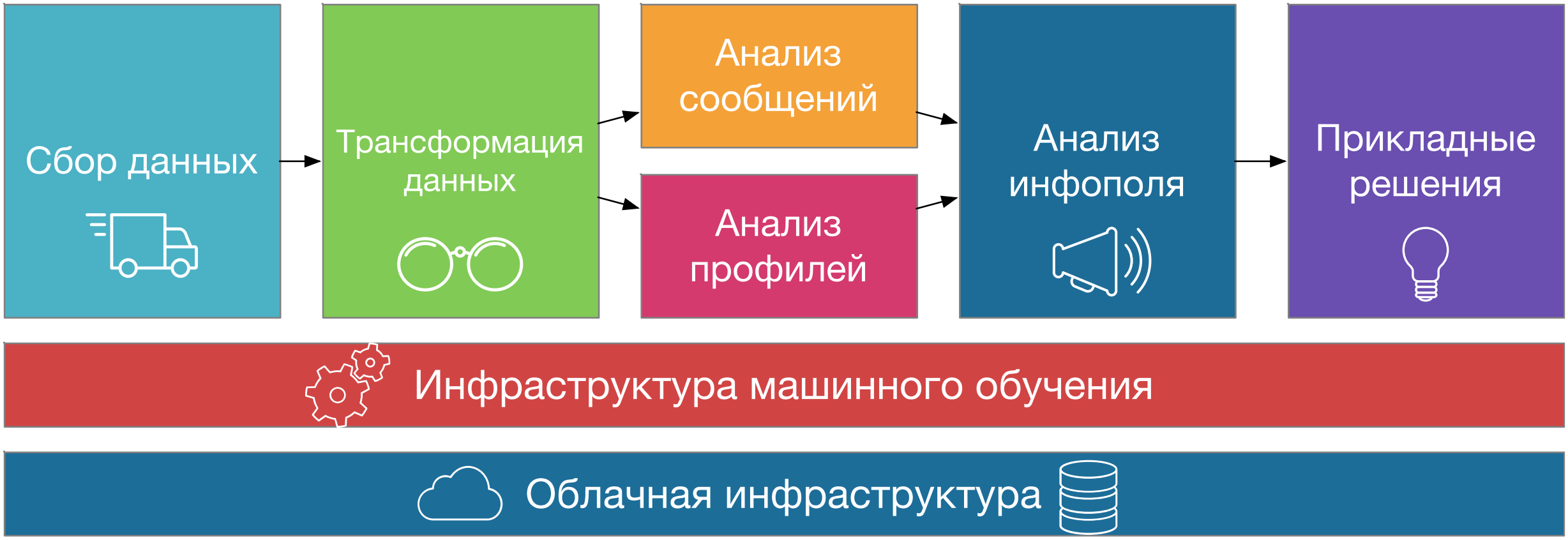
- Модель информационной компании в социальных медиа
- Получение репрезентативной выборки сообщений

Решения

- Разработана *формальная модель* развития информационной кампании
- Выделены *признаки* и реализованы методы их обнаружения (резкие всплески негативных высказываний, дубликаты сообщений, о целевых объектах, в различных ресурсах от разных авторов, использование ботов и др.)
- Разработаны *методы сбора* данных из СМИ (3000 источников), социальных сетей (~2 млн. групп ВК), блогов (ЖЖ) и др.
- Разработана распределённая система для обнаружения информационных кампаний на ранних стадиях (не более 1,5 часов после появления) на основе Apache Spark и Apache Ignite.
- Обработка ~ 4Гб текстовых сообщений в день

Турдаков Д. Обнаружение информационных кампаний в социальных медиа // 10я Всероссийская межведомственная научная конференция “Актуальные направления развития систем охраны, специальной связи и информации для нужд органов государственной власти Российской Федерации”, Орел, 2017.

Платформа Talisman. Этапы обработки данных



Нефункциональные требования к платформе

Эффективность

- Минимизация скорости отклика
- Масштабируемость, скорость обработки данных
- Интеграция со сторонними системами
- Минимизация использования физических ресурсов

Продуктивность

- **Использования**
  - Точность и полнота ответов
  - Удобные интерфейсы (HCI)
- **Разработки**
  - Гибкость при добавлении новых функций
  - Перенос на новую предметную область

Безопасность

- Соответствие требованиям к разработке безопасного ПО
- Резервирование, синхронизация данных
- Авторизация, аутентификация
- Разделение прав и контроль доступа, журналирование
- Защита от внешних вторжений и минимизация возможности утечки чувствительных данных, обеспечение конфиденциальности потребителя (программно-аппаратная архитектура)
- Независимость от вендора
- Отчуждаемость

Сбор данных

Вызовы

1. Обеспечение заданной **полноты** и точности в условиях ограниченных ресурсов
2. Отсутствие фиксированных программных интерфейсов
3. Обеспечение актуальности данных
4. Защита от сбора данных
5. Обработка динамических страниц

Решения

- Построение моделей социальных сетей для доказательства свойств алгоритмов сбора данных (точность, полнота при заданных ограничениях на число запросов)
- Собственная инфраструктура сбора открытых данных
- Ежедневный мониторинг более 2 миллионов источников
  - социальных сетей (VK, Facebook, Instagram, Twitter)
  - блогов (LiveJournal)
  - сайтов СМИ
  - новостных агрегаторов (Яндекс.Новости)
  - Форумы Dark web
- Интегрированные средства визуальной разметки пользовательских интерфейсов

1. Varlamov M., Turdakov D. **A survey of methods for the extraction of information from Web resources** // Programming and Computer Software. — 2016.
2. Yatskov A., Varlamov M., Turdakov D. Y. **Extraction of Data from Mass Media Web Sites** // Programming and Computer Software. — 2018
3. M. Drobyshevskiy, D. Aivazov, D. Turdakov, A. Yatskov, M. Varlamov, D. Shayhelislamov. **Collecting Influencers: a Comparative Study of Online Network Crawlers** // Proceedings of ISPRAS open conference. — IEEE. 2019.

Трансформация данных

Вызовы

- Поиск неявных дубликатов в данных
- Извлечение информации из полустурктурированных данных (таблиц, pdf документов и др.)
- Обработка изображений
  - OCR: фотографии документов при сложном освещении под углом
  - Быстрый поиск по лицам в динамически меняющейся базе

Решения

- Автоматизированные методы дедупликации данных, на основе моделирования действий эксперта
- Методы и программные средства извлечения информации из полуструктурирвоанных данных
- TESSERACT + пред- и пост- обработка
- Разработка knn индексов

1. Y. Nedumov, D. Turdakov, V. Maiorov, P. Ovchinnikov. Automation of data normalization for implementing master data management systems // Programming and Computer Software. — 2013.
2. Astrakhantsev N., Turdakov D., Vassilieva N. Semi-automatic Data Extraction from Tables. // Proceedings of the SYRCODIS 2013 Colloquium on Databases and Information Systems. — 2013
3. Manuk Akopyan, Oksana Belyaeva, Timofei Plechov, Denis Turdakov, Text Recognition on Images from Social Media. 2019 Ivannikov Memorial Workshop (IVMEM)

Анализ сообщений

Вызовы

- Работа на уровне понятий, а не ключевых слов
- Связывание сущностей (entity linkage)
  - Разрешение многозначности
  - Синонимия
- Анализ эмоциональной окраски сообщений

Решения

Текстерра (texterra.ispras.ru)

система автоматического построения онтологий и семантического

- **Один из наиболее быстрых инструментов обработки текстов (**82Kb/sec полный семантический анализ)
- Поддержка языков: **английский, русский**, корейский*, армянский* (идентификация >60 языков, морфология 56)
- Собственная быстрая система управления базами знаний
- Свободный масштабируемый программный интерфейс (API)
- Более 20 инструментов анализа текстов с точностью и полнотой мировых аналогов (постоянно внедряются последние достижения области):
- *Распознавание именованных сущностей (F1: 0.77-0.87), разрешение лексической многозначности (Acc: 0.76), анализ эмоциональной окраски (F1: 0.74-0.98) и др.*

1. Texterra: A framework for text analysis / D. Turdakov, N. Astrakhantsev, Y. Nedumov, A. Sysoev, I. Andrianov, V. Mayorov, D. Fedorenko, A. Korshunov, S. Kuznetsov // Programming and Computer Software. — 2014.
2. High Precision Method for Aspect Extraction in Russian. Computational Linguistics and Intellectual Technologies / V. Mayorov, I. Andrianov, N. Astrakhantsev, V. Avanesov, I. Kozlov, D. Turdakov. — 2015.
3. Astrakhantsev N., Fedorenko D., Turdakov D. Automatic enrichment of informal ontology by analyzing a domain-specific text collection // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue. — Springer, 2014

Анализ профилей

Вызовы

- В дискуссиях участвуют боты, которые притворяются реальными людьми
- Пользователи не указывают информацию о себе или указывают неверно
- Локальной информации о профиле часто недостаточно для корректного восстановления значений атрибутов

Решения

- Распределенный подсчет векторных представлений вершин (graph embedding) для графов со 100+ млн. вершин и 10+ млрд. ребер.
- Разработаны методы определение стационарных демографических атрибутов *(пол, возраст, регион проживания, семейное положение, образование, род деятельности,* F_1 ~80%)
- Разработан метод выявления ботов (F_1– 65%-82%)
- Сопоставление профилей в разных сетях (F_1 - 89% при сопоставлении контактов вершины)

1. Trofimovich J., Kozlov I., Turdakov D. User location detection based on social graph analysis // Proceedings of ISPRAS open conference. — 2016.
2. Skorniakov K., Turdakov D., Zhabotinsky A. Make Social Networks Clean Again: Graph Embedding and Stacking Classifiers for Bot Detection // Proceedings of the 2nd International Workshop on Rumours and Deception in Social Media. — 2018.
3. A. Gomzin, A. Laguta, V. Stroev, D. Turdakov. Detection of author's educational level and age based on comments analysis // Dialogue. — 2018.

Анализ информационного поля

Вызовы

1. Локальная информация не дает общей картины
2. Множество игроков определяют информационную повестку (на профессиональной основе)
3. Для своевременной реакции необходимо понимать тенденции на этапе их зарождения

Решения

1. Выявление информационных сюжетов (потоковая кластеризация с учетом специфики языка сообщений)
2. Моделирование потоков распространения информации для предсказания популярности информационных сюжетов на ранних этапах
3. Определение ролей пользователей
4. Определение точки зрения пользователя в дискуссии

1. Avetisyan A., Drobyshevskiy M., Turdakov D. Predicting the Popularity of News Stories in the Early Stages of Dissemination // Proceedings of ISPRAS open conference. — 2019.
2. Турдаков Д. Обнаружение информационных кампаний в социальных медиа // 10я Всероссийская межведомственная научная конференция “Актуальные направления развития систем охраны, специальной связи и информации для нужд органов государственной власти Российской Федерации”, Орел, 7-8 февраля. — 2017.

Инфраструктура машинного обучения

Вызовы

1. Минимизация усилий на получение размеченных данных для обучения моделей
2. Проблема устаревания моделей машинного обучения (concept drift, feature drift) в работающей системе
3. Объяснение моделей (Explainable machine learning)

Решения

- Разметка данных
  - Специальные приложения для разметка одиночных объектов (бот telegram) и разметки последовательностей
  - Активное обучение, краудсорсинг
- Проблема устаревания моделей
  - Проактивное обучение
  - Адаптация к новой предметной области
  - Автоматическое дообучение и контроль версий моделей
- Объяснение моделей
  - Сочетание методов основанных на словарях со сложными классификаторами (обучение с частичным привлечением учителя)

1. Рындин М., Турдаков Д. Проактивная разметка примеров для адаптации к домену // Proceedings of ISPRAS open conference. — 2019.
2. Gilyazev R., Turdakov D. Active Learning and Crowdsourcing: A Survey of Optimization Methods for Data Labeling // Programming and Computer Software. — 2018.
3. Bartunov S., Tutdakov D. Wikifyme: Creating testbed for Wikifiers // Труды Института системного программирования РАН. — 2011. — Т. 21.

Облачная инфраструктура

Вызовы

1. Необходимость в большом количестве ресурсов для пакетной обработки данных и обучении моделей
   - Периодическая
   - Кратковременная
2. Постоянно держать ресурсы в резерве для задачи - дорого

Решения

Облачная инфраструктура на основе OpenStack и Kubernetes

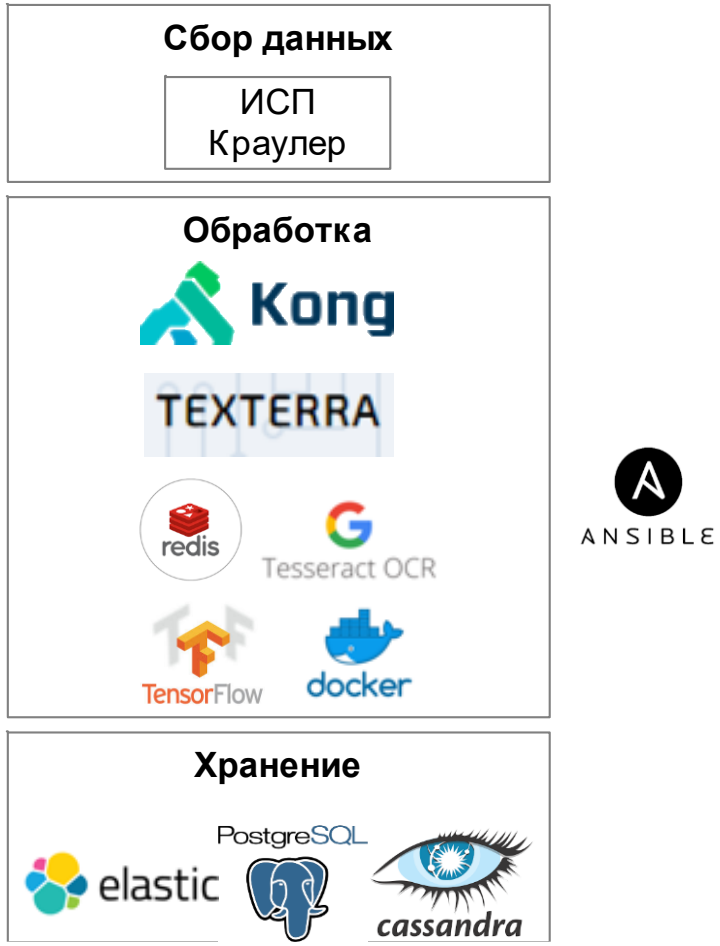
- Облачная среда Асперитас с упрощенным развертыванием в локальном окружении
- Инструменты для установки (оркестрации) систем распределенной обработки больших данных (HDFS, Apach Spark, Cassandra и др.)

1. Oleg Borisenko, David Badalyan. Evaluation of SQL benchmark for distributed in-memory Database Management Systems IJCSNS International Journal of Computer Science and Network Security, VOL.18 No.10, October 2018
2. O.D. Borisenko, N.A. Lazarev Implementing JSON operations for In-memory Data Grid as passthrough cache layer to RDBMS International Journal of Civil Engineering and Technology 9(10):1033-1040 · October 2018
3. Аксенова Е.Л., Швецова В.В, Борисенко О.Д., Богомолов И.В. Реализация сервиса для замены Keystone в качестве центрального сервиса идентификации облачной платформы Openstack Труды Института системного программирования РАН. Том 29, выпуск 6, 2017 г. Стр. 203-212.

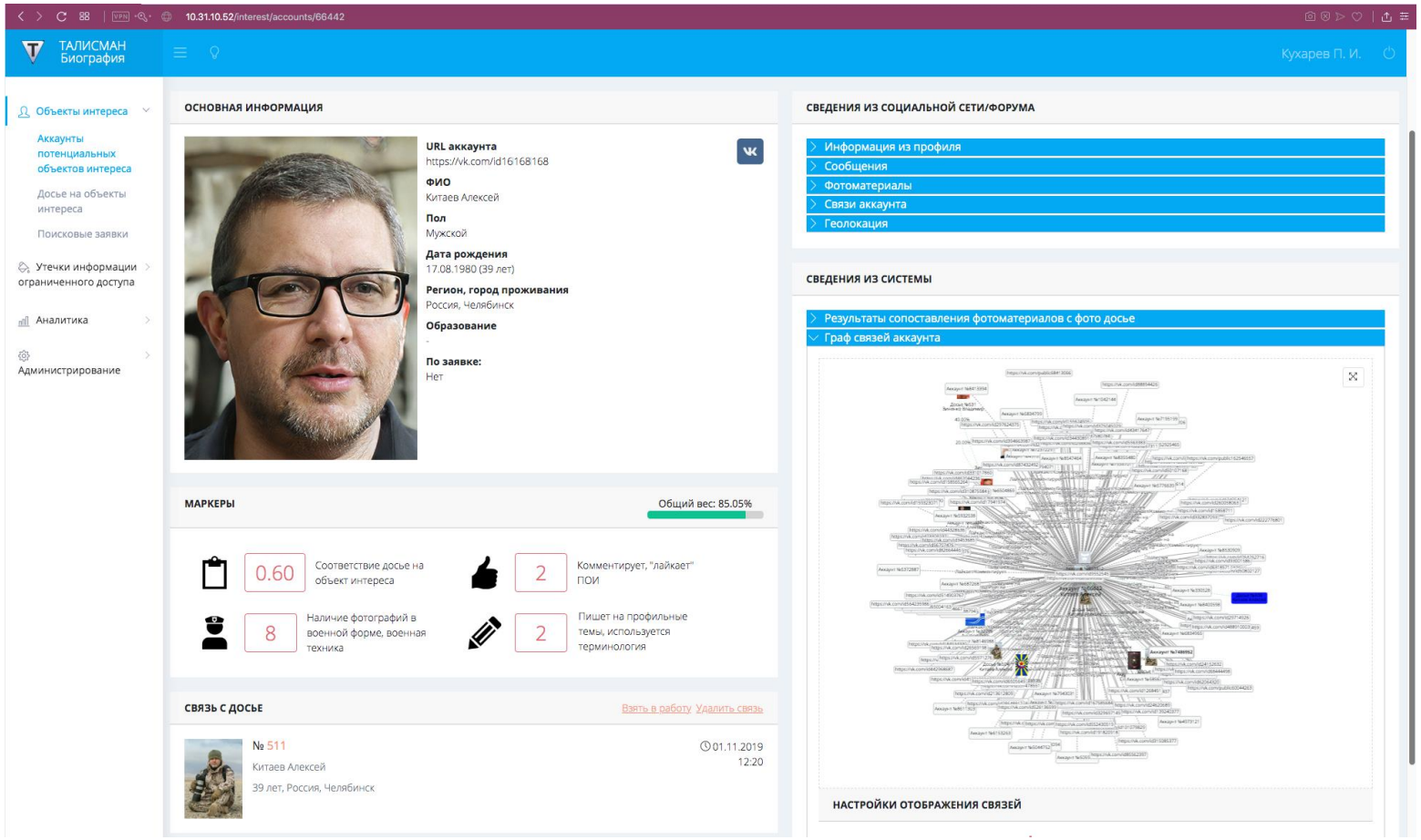
Talisman.Поток

Область применения

- Система для предобработки потока больших данных из социальных медиа
- Масштабируемая модульная система на основе контейнеров (Docker) и шлюза программных интерфейсов (Kong)
- Повышает продуктивность разработки прикладных систем анализа потоковых данных
- Отличается высокой скоростью обработки информации



Talisman.Биография



Область применения

- Задачи отдела кадров
- Задачи отдела по связям с общественностью
- Связывание данных сотрудников или соискателей с их аккаунтами в социальных сетях
- Верификация анкетных данных
- Обнаружение утечки корпоративной информации через аккаунты сотрудников
- Управление репутацией организации

ИСП РАН

Фундаментальные результаты

Анализа социальных медиа

- Масштабируемые методы обнаружения пересекающихся сообществ
- Модели генерации графов со свойствами реальных социальных сетей
- Методы предсказания значений стационарных демографических атрибутов пользователей
- Масштабируемый метод построения векторного представления графа
- Метод выявления ботов
- Методы выявления информационных кампаний

Интеллектуальный анализ текстов

- Методы автоматизированного построения онтологий предметной области
- Методы анализа неформальных текстов и их привязки к объектам онтологии (разрешение многозначности)

Машинное обучение

- Архитектура системы и методики, обеспечивающие решение как проблем разметки данных для обучения, так и поддержания обученных моделей в актуальном состоянии

Полученные результаты реализованы в программных технологиях и доведены до промышленного использования

- D. Y. Turdakov, Y. R. Nedumov, and A. A. Sysoev, *Method to build a document semantic model*. Патент РФ № 2011148742 от 30.11.2011. Патент США № US20130138696A1 от 30.05.2013
- Бартунов, С. О. and Коршунов, А. В. and Турдаков, Д. Ю. and Кузюрин, Н.Н. and ПАРК Сеунг-Таек and РЫУ Вонхо and ЛИ Хыунгдонг. Способ интеграции профилей пользователей онлайновых социальных сетей. Патент РФ №2469389 от 08.11.2011
- Лизоркин, Д. А. and Гринев, М. Н. and Велихов, П. Е. and Турдаков, Д. Ю. Итерационный способ получения функции похожести между объектами со ссылками. Патент РФ № 2413291 от 25.02.2009
- Ким. Ханг-Киу, Чихрадзе. Кирило. Константинович, Коршунов. Антон. Викторович, Пастухов. Роман. Константинович, and Турдаков. Денис. Юрьевич, *Способ и устройства для распределенной генерации случайных социальных графов со структурой пересекающихся сообществ пользователей*. Патент РФ, RU2014117945A №2014117945/08 от 05.05.2014

ИСП РАН

Спасибо за внимание!