

# Обработка естественного языка и анализ социальных сетей в условиях взрывного роста данных

к. ф.-м. н. Турдаков Денис Юрьевич

# ОБЩИЙ ВЗГЛЯД



# Технологии ИСП РАН



Native XML Database System

- XML СУБД с открытым кодом
- Международное сообщество пользователей и разработчиков
- Более 100 загрузок ежедневно

# Сервисы ИСПРАН

ИСПРАН предоставляет более 25 сервисов для обработки естественного языка, анализа социальных сетей и работы с базами знаний

ISPRAS API provides more than 25 tools based on natural language processing, social networks analysis, and knowledge base utilization

Get your API key

### Tools

- **Text processing Tools** take any text as input and returns enriched content with meaningful annotations
  - **Standard NLP Tools** include tokenization, part of speech tagging, NERC and so on
  - **KB-based NLP Tools** detect terms, disambiguate meanings, found key ones by using semantic from our Knowledge base
  - **Sentiment analysis Tools** analyze sentiments of text's author with respect to domain and several aspects
- **Twitter NLP Tools** identify demographic attributes of any Twitter user.
- **Direct queries to Knowledge base** provide full access to our Knowledge base: terms, meanings, relations between them, their specific attributes

### Demos

Demonstrations of eminent capabilities of these tools

- Text Processing Demo**  
suggests to enter any text and get information about its sentiment, or terms and concepts, or only key ones
- Twitter Demo**  
finds all demographic attributes of any Twitter user by his/her Twitter name or tweets
- VizOntia**  
visualizes our Knowledge base in user-friendly manner - explore graph or terms and concepts

### Solutions

End-point applications based on these tools

- ReputationNoon**  
monitors reputation of any person, product, or any other entity, including several ones at time
- TVNoon**  
suggests handy tools to explore movies, considering granular genres, user reviews/ratings, and special aspects like plot or soundtrack
- BookOnMap**  
transforms any book into the fascinating trip over all places mentioned in the book with exhaustive descriptions from Wikipedia

<https://api.ispras.ru/>

# Обработка естественного языка

# Приложения обработки текстов

- Семантический поиск
- Автоматическое реферирование
- Анализ эмоциональной окраски сообщений
- Определение демографических атрибутов пользователей Интернета
- ...

# Текущее состояние области

- Коммерческие сервисы



- Библиотеки и системы с открытым кодом



- Решаются задачи уровня морфологии и синтаксиса
- Для решений уровня семантики нужны знания о мире в понятном для компьютера виде
- Построение баз знаний вручную чрезвычайно трудоемко и дорого

# Источники знаний



- Источники знаний о многих предметных областях
- Вызов: Специализированные области не покрыты

# Технология Texterra

Texterra - масштабируемое решение для быстрой обработки текстов, основанное на использовании знаний, извлекаемых из Веб-ресурсов и коллекций текстовых документов

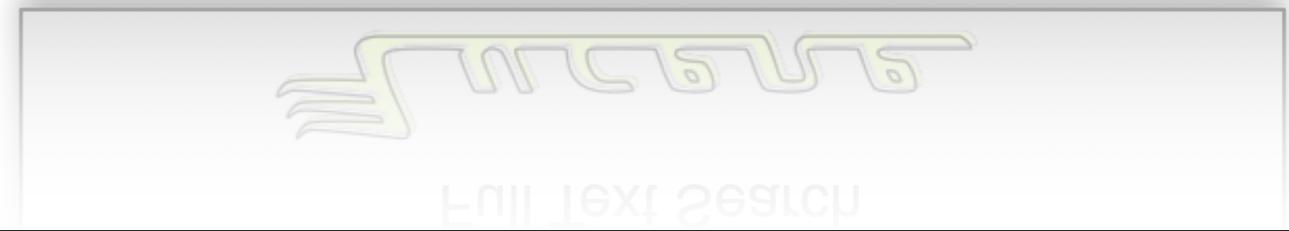
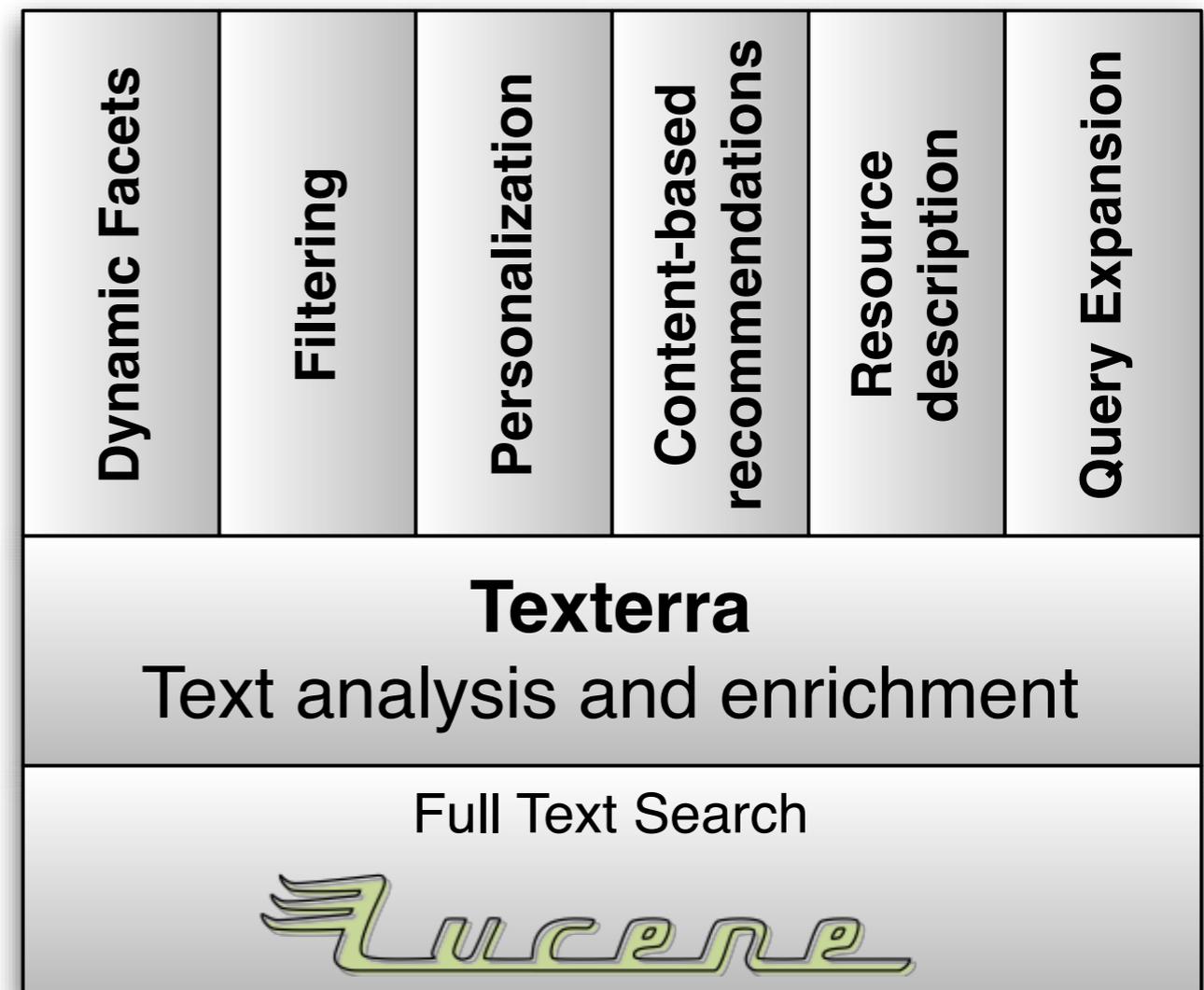
# Преимущества

**Используется подход аналогичный наиболее известным инструментам, но дополнительно предоставляется**

- Инструменты для автоматического
  - построения базы знаний, используя открытые Веб-ресурсы
  - обогащения базы знаний предметно-ориентированной информацией на основе анализа текстовых документов
- Использование знаний о мире позволяет получить более точные результаты при решении прикладных задач
- Многоязычность: поддержка русского, английского, корейского + возможность добавления новых языков
- Высокая скорость работы

# Технология \*Noon

Набор инструментов для быстрой разработки предметно-ориентированных семантических поисковых и навигационных систем



# \*Noon: поиск

- Texterra преобразует поиск по ключевым словам в семантический поиск по значениям
- Эта операция улучшает точность ранжирования результатов поиска

# \*Noon: навигация

- Что делать если ключевые слова для запроса неизвестны?
- \*Noon предоставляет инструменты для навигации по коллекциям документов
- Аналог навигации по карте в Веб-сервисах картографии ([maps.yandex.ru](https://maps.yandex.ru))
- Позволяет исследовать новую область и понять ключевые слова для последующего поиска

# **Анализ социальных сетей**

# Решения ИСП РАН

Анализ сетевой структуры

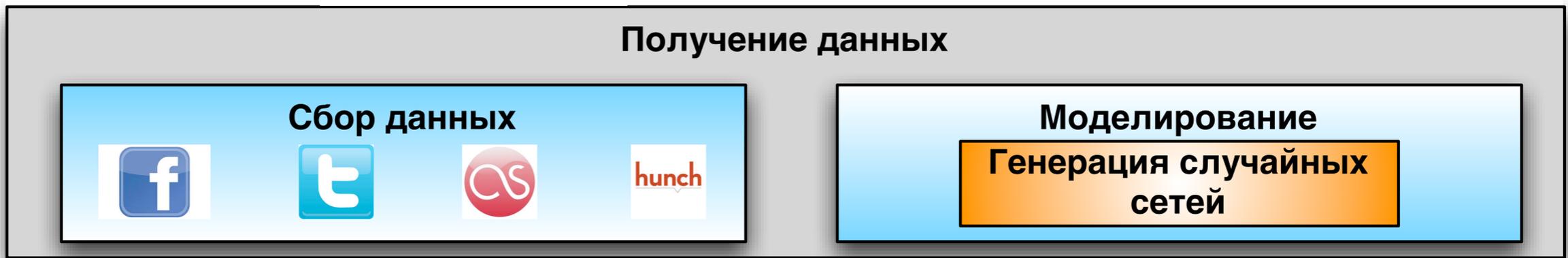
Анализ текстов



Масштабируемые системы для обработки сетевых данных



Получение данных



## Twitter Processing Demo

Twitter Link Custom Info

BillGates

Autodetect language ▾

Submit

## Result

User is single male republican over 23 with higher education who believes in god.

## Description

User Language: English

## Extracted Attributes

- **Age:** over 23
- **Gender:** male
- **Relationship:** single
- **Politics:** republican
- **Education:** higher education
- **Religious:** yes

Поддерживается  
8 языков

# Извлечение демографических атрибутов

## User Info



@BillGates

Bill Gates

Sharing things I'm learning through  
my foundation work and other  
interests...

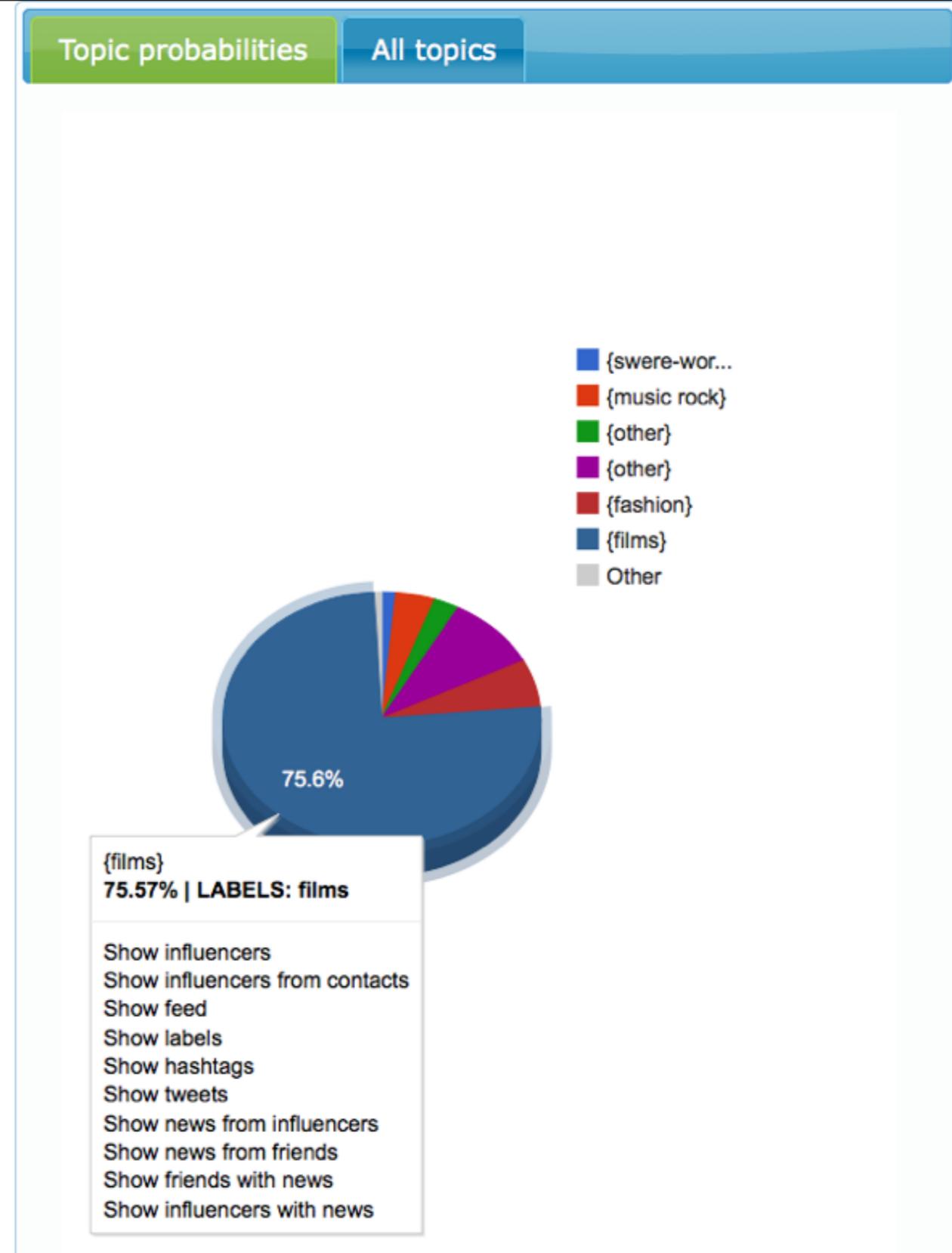
## Example Tweets

- 8 out of 10 people say they want to give to a charity. #GivingTuesday makes it easy: <http://t.co/a0LGvYyEPw> <http://t.co/VOFZYc2uzb>
- Progress towards the MDGs has been impressive. Are we ready to take the next step? @JeffDSachs weighs in: <http://t.co/LZzin2by6G>
- When you see the term "global health," think "saving lives." That's what it means. | @nytimes: <http://t.co/1YPmfasp9J>
- The idea that aid is wasted is just wrong. I recently spoke with the Sun newspaper about the impact of the UK's aid: <http://t.co/KH8tTWLBLO>
- Here's my wrap-up of day 2 in NYC. Got to see leaders from #Chad, #Pakistan, and much more: <http://t.co/HXXcLiP9S6>

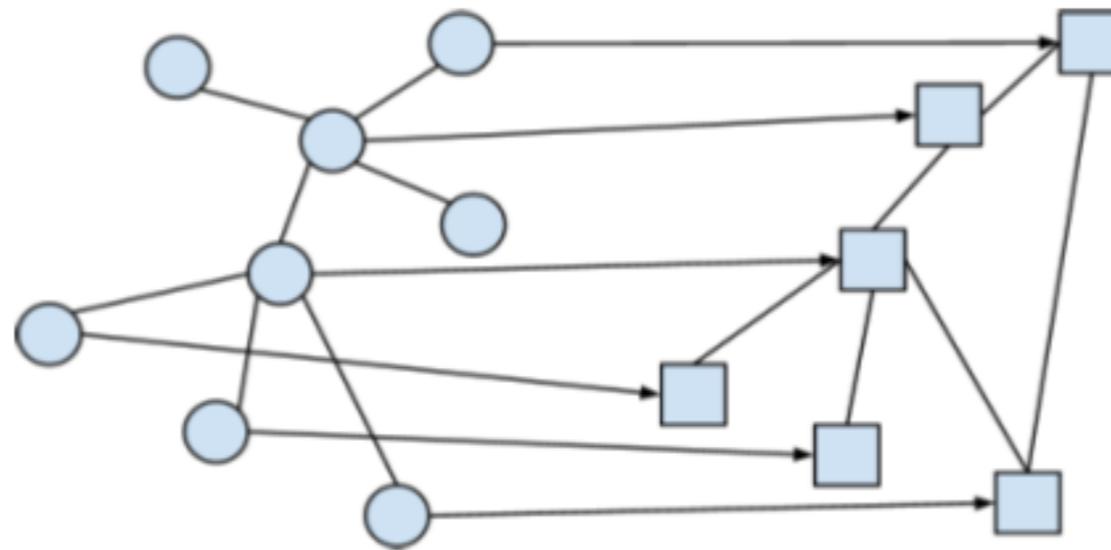
- Пол
- Возраст
- Семейное положение
- Политические взгляды
- Образование
- Религиозные взгляды

# Поиск агентов влияния

- Граф влияния в Twitter с привязкой по темам
- На основе распределенных алгоритмов тематического моделирования
  - PLSA
  - robust PLSA



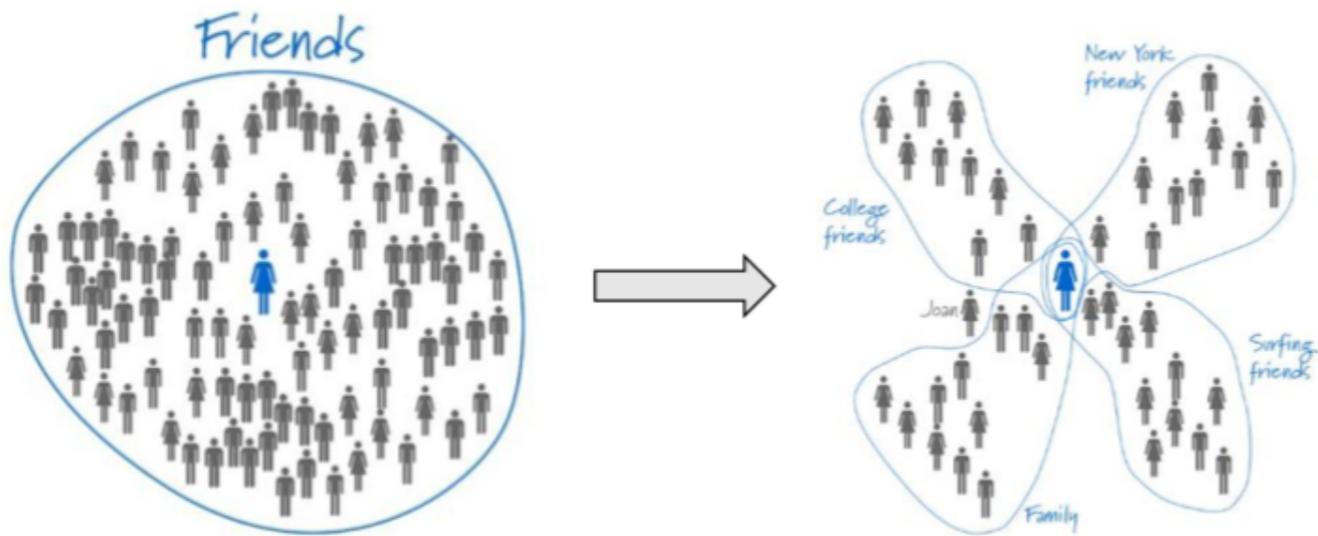
# Идентификация ПОЛЬЗОВАТЕЛЕЙ



- Сопоставление аккаунтов пользователей в разных социальных сетях, используя
  - текстовые данные
  - структуру сети
- Позволяет производить деанонимизацию

# Поиск сообществ

Локальные  
сообщества



- Автоматическое структурирование контактов
- Улучшение рекомендательных систем

Глобальные  
сообщества



- Оптимизация трафика
- Улучшение рекомендательных систем
- Поиска спама

Обработка x100 млн. пользователей!

# Обработка “больших данных”

- Необходимы технологии для обработки больших объемов данных
- Использование систем с открытым кодом позволяет решать задачи уже сейчас



# Участие ИСП РАН

- Во время решения практических задач возникают новые требования к открытым системам
- ИСП РАН активно участвует в развитии международных проектов
- Разработанные дополнения и модификации признаны сообществом

