ИТМО
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Сетецентрические технологии сбора данных в Интернет

Якушев А. В., НИИ НКТ
Дейкстра Л., НИИ НКТ

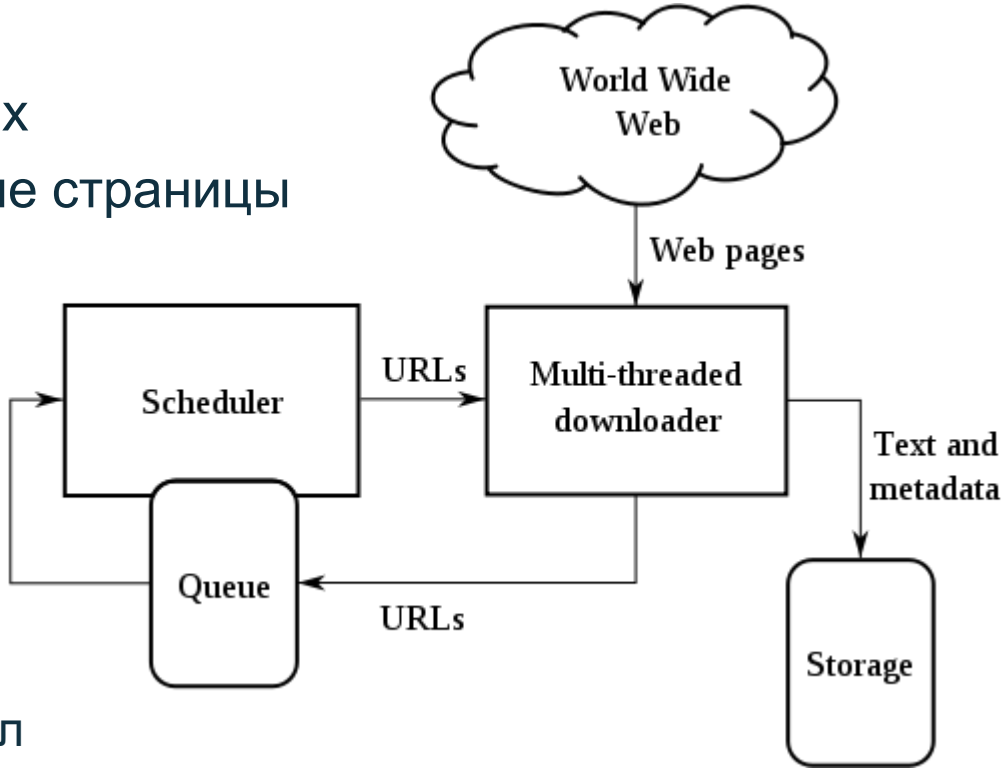
Санкт-Петербург, 11 Октября 2012

Сетецентрические технологии

- ✓ Что такое сетецентрические технологии
  - Технологии в которой активно используется знание о сети и связях
- ✓ Сети – они везде
  - Биологические: хищник-жертва, гены в ДНК, …
  - Компьютерные сети: соединения компьютеров, ..
  - Социальные сети: цитирования, упоминания, Интернет
- ✓ Краулер – инструмент для сбора данных из Интернета
  - Real-time краулеры
    - Быстрый доступ к собранным данным
    - Систем мониторинга
    - Операции осуществляются в памяти
    - Сложность архитектуры и стоимость поддержки
  - Batch краулеры
    - Обработка данных больших объемов
    - Подходит и для систем мониторинга, но не таких быстрых
    - Использует эффективные алгоритмы работы с диском

Краулер - Технология сбора данных

- ✓ Алгоритм
  - Скачиваем данные
  - Сохраняем в базу данных
  - Находим ссылки на новые страницы
  - Добавляем их в очередь
  - Сортируем и фильтруем
    - Политика обхода
  - Скачиваем
- ✓ Политики обхода
  - Учет структуры сети
    - Ссылки ведущие на узел
    - Важность узла - PageRank



Краулеры для социальных сетей

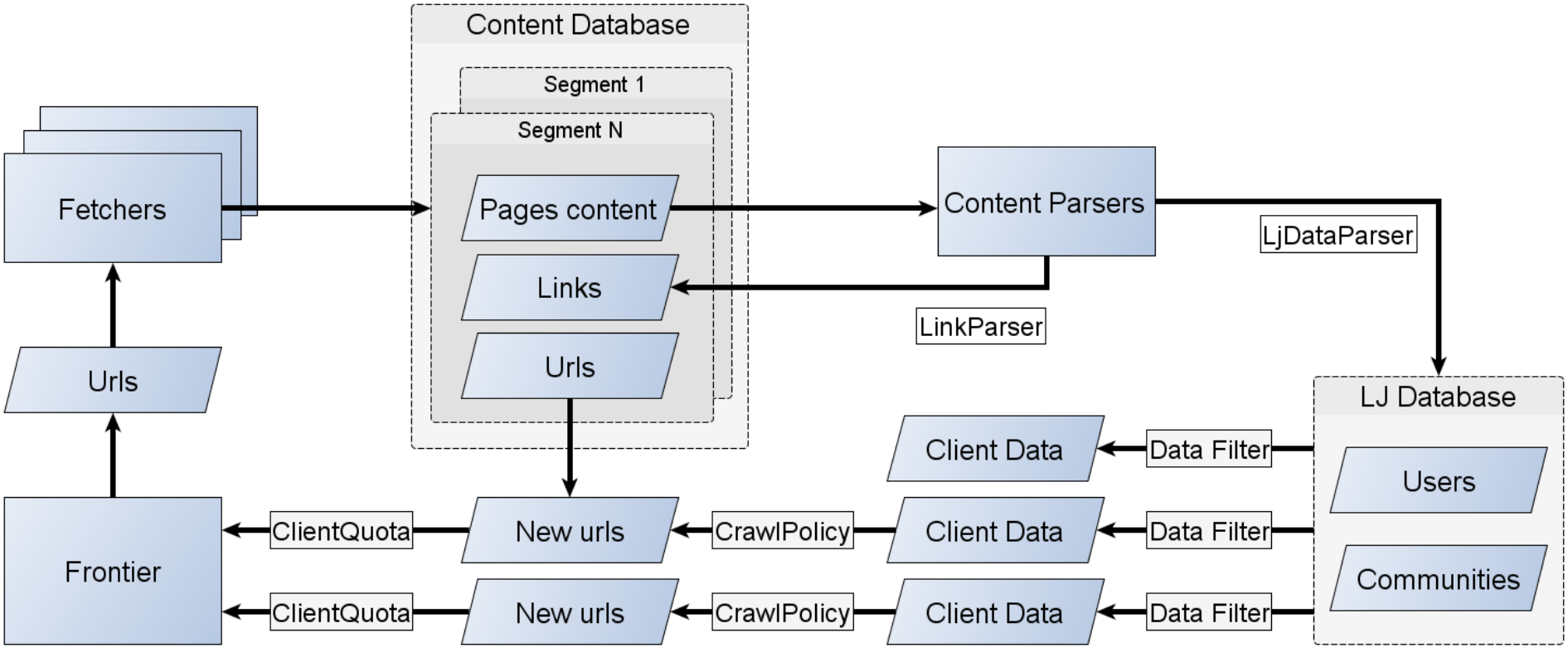
- ✓ Сбор данных о пользователях
  - Работа со структурированными данными
    - Атрибуты пользователей, связи между ними,..
- ✓ Нужны дополнительные уровни в архитектуре
  - Выделение атрибутов пользователей/сообществ
  - Объединение данных в единый «контекст» пользователя
- ✓ Ограничения на число запросов к социальной сети
- ✓ Более эффективные политики обхода
  - Более полная информация о пользователе
  - Заранее можем принимать решения о важности пользователя

Данные о пользователе

- ✓ Атрибуты пользователя
  - Личная информация, интересы
  - Тексты: большие и малые
    - Из больших мы можем извлекать информацию
    - Малые – анализ тональности текстов
- ✓ Ссылочные данные
  - Сеть «дружбы» пользователей
  - Упоминания пользователей друг друга
- ✓ Оценки важности данных
  - Репосты, retweets
  - Число Like’ов
- ✓ Политика обхода на основе этих данных
  - «Рукописные правила» – на основе числа ссылок
  - Машинное обучение

Архитектура нашего краулера

- ✓ Распределенный batch-краулер
  - MapReduce – как основа для организации распределенных вычислений
    - Эффективная работа с сотнями гигабайт данных
    - Устойчивая к различным ошибкам работа
- ✓ Работа с неструктурированными данными из Интернета и структурированными из социальных сетей
- ✓ «Многопользовательский» режим

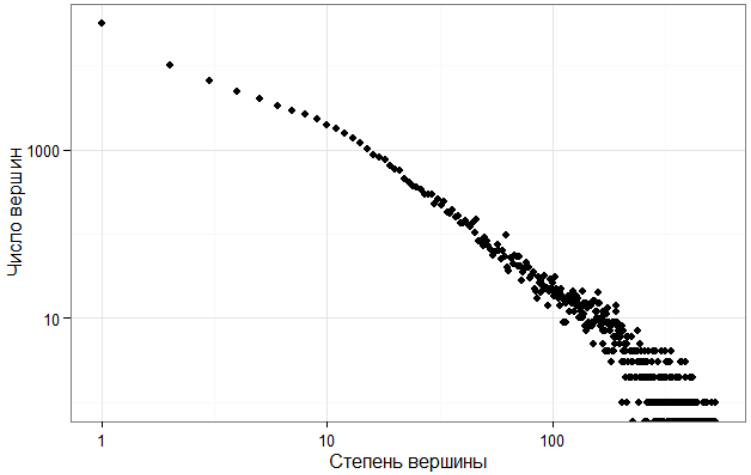


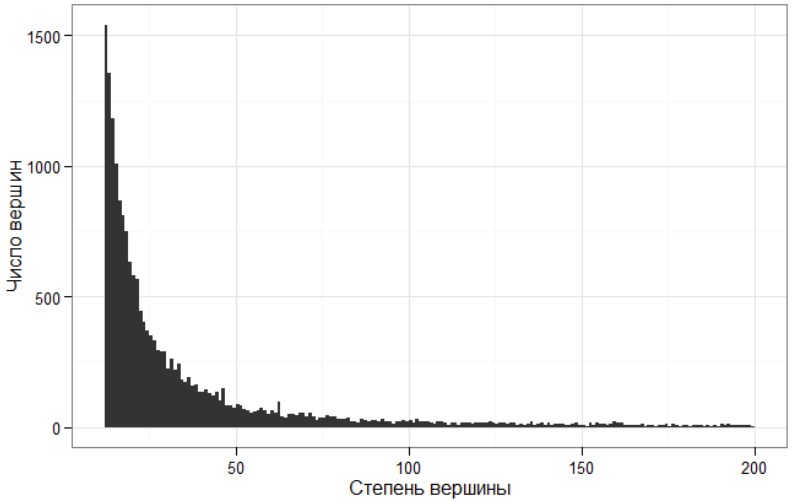
Политика обхода и фильтрация пользователей

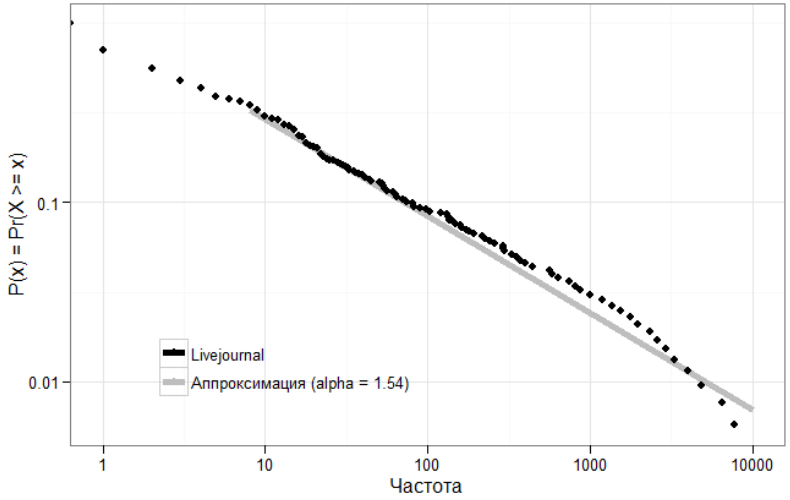
- ✓ Политика обхода
  - Приоритет пользователя - на основе ссылочной информации
- ✓ Фильтруем пользователей, анализируя их тексты
- ✓ Классификация текстов
  - Нету обучающей коллекции
  - Используем словарь взвешенных терминов, описывающих предметную область
  - Считаем вес документа и сравниваем с пороговой величиной
  - Другие функции ранжирования - Okapi BM25

Исследование сетей в Livejournal

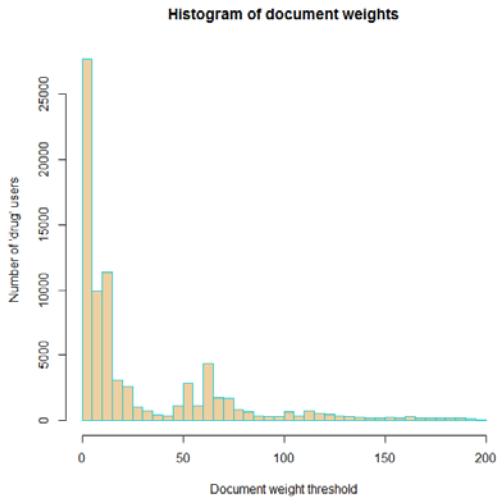
- ✓ Сети дружбы: «друзья», «в друзьях у», «взаимные друзья»
- ✓ Сеть «упоминаний»
  - ссылка на пользователя в тексте
- ✓ Все сети - scale-free: $p(x) = Cx^{-\alpha}$
  - «легко» разрушаемы

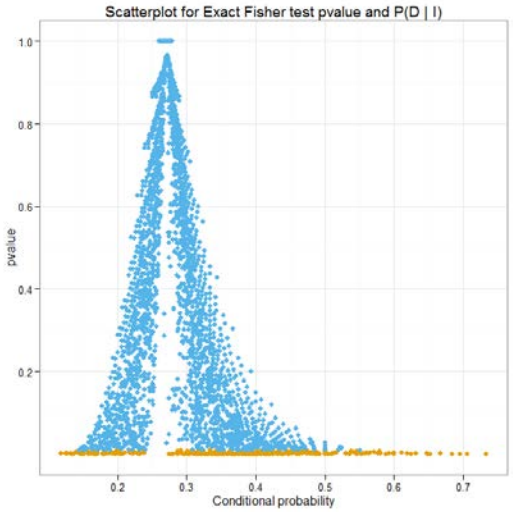


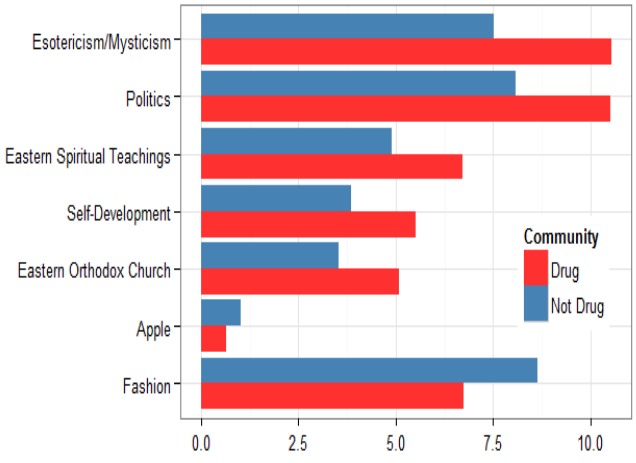


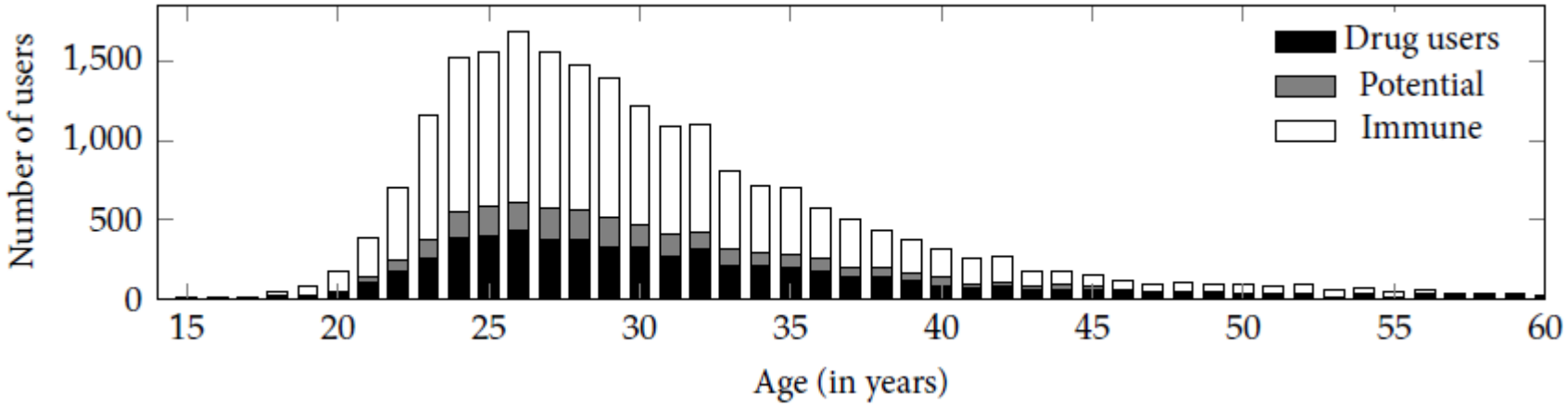


Исследование пользователей, пишущих о наркотиках









Выводы

- ✓ Связи между пользователями – это
  - Пути распространения информации
  - Возможности влиять на людей
  - Позволяют находить связи между «атрибутами» пользователей
- ✓ Сети позволяют:
  - Выявлять группы скрытых пользователей, общающихся преимущественно друг с другом
  - Определять роли элементы в процессе распространения информации и находить ключевые элементы
- ✓ А если объединить сети с данными о пользователях, то можно:
  - Выявлять факторы влияющие на образование новых связей
  - Создавать более качественные рекомендательные системы, учитывающие контекст пользователя и его связи друг с другом

Вопросы?