

Информационный поиск

Лукашевич Наталья Валентиновна
Ведущий научный сотрудник НИВЦ МГУ,
профессор ВМК МГУ,
профессор филологического факультета МГУ

Тихомиров Михаил Михайлович
(младший научный сотрудник НИВЦ МГУ)

vmk_ir@mail.ru

Поиск информации и информационный поиск (Information Retrieval)

- Поиск информации в Интернет – это повседневная деятельность многих людей
- Поиск информации и общение – это наиболее популярные виды использования компьютеров
- Приложения, использующие поиск информации, - везде вокруг нас
- Сфера науки, которая исследует методы поиска информации, называется информационный поиск (*information retrieval (IR)*)
 - искать можно разные виды информации
 - основной фокус информационного поиска с 50-х годов – на тексты и документы

Что такое документ?

- Примеры
 - Интернет-страницы, электронные письма, книги, новости, посты форумов, патенты и многое другое
- Общие свойства
 - Значительное текстовое содержание
 - Некоторая структура:
 - заголовок, автор, дата - для статей;
 - тема, отправитель, адресат - для писем

Документы vs. записи базы данных

- Записи базы данных (структурированные таблицы) состоят из хорошо определенных полей и атрибутов
 - е.г., банковские записи балансы, номера счетов, имена, адреса, даты рождения, номера социального обеспечения
- Легко сопоставлять запросы и поля таких баз данных (хорошо определенная семантика)
- Текст более сложный – неструктурированная информация

Документы и записи базы данных

- Запрос к базе данных
 - *Найти записи с суммами, большими \$50 тысяч рублей*
 - Легко сопоставить со значением определенного поля в базе данных – структурированная информация
- Пример запроса
 - *Банковские скандалы в США*
 - Этот запрос нужно сравнивать с неструктурированными текстами новостей

Сравнение текстов

- Сопоставление текста запроса с текстом документа и определение того, что такое хорошее сопоставление – базовый вопрос информационного поиска
- Точное сопоставление слов - недостаточно
 - Много различных способов сказать одно и то же на естественном языке
 - е.g., преступность в Сибири
 - Некоторые документы подходят к запросу лучше, чем другие

Измерения информационного поиска

- Информационный поиск – это больше чем поиск по текстам, и больше, чем просто интернет-поиск
 - Хотя эти вопросы являются центральными!
- Поиск осуществляется на основе разных типов данных, разных типов приложений и разных задач

Другие исходные данные (нетексты)

- Поиск по нетекстовым данным
 - видео, фото, музыка, речь
- Их содержание также трудно описывать и сравнивать
 - Текст может использоваться для описания (теги)
- Подходы, созданные для классического информационного поиска являются приемлимыми и для нетекстовых данных

Задачи, связанные с поиском информации

- Ad-hoc поиск
 - Найти релевантный документ в ответ на произвольный запрос
- Фильтрация
 - Отобразить нужные пользователю документы
- Классификация
 - Проставить рубрики документам
- Ответы на вопрос
 - Дать ответ на заданный вопрос
- Визуализация выдаваемой информации
 - Аннотации (рефераты) и др.

Измерения информационного поиска

Содержание	Приложения	Задачи
Текста	Веб-поиск	Поиск по запросу
Картинки	Предметно-ориент. поиск	Фильтрация
Видео	Корпоративный поиск	Классификация
Сканы	Десктопный поиск	Ответы на вопросы
Аудио	Поиск на форумах	
Музыка	Поиск патентов	

Важные понятия в информационном поиске

- Relevance – релевантность
- Evaluation - оценка качества
- Users and Information Needs –
потребность пользователя,
информационная потребность

Релевантность

- Что это?
- Простое (и упрощающее) определение:
Релевантный документ содержит информацию, которую искал пользователь, когда задавал запрос поисковой машине
- На релевантность оказывают влияние много различных факторов: задача, контекст, опыт пользователя, новизна, стиль
- Тематическая релевантность (отражение заданной темы) vs. пользовательская релевантность (все остальные факторы)

Релевантность и модели поиска

- Модели поиска отражают «взгляд» на релевантность
- Ранжирующие алгоритмы, используемые в поисковых машинах базируются на моделях поиска
- Большинство моделей описывают статистические свойства текстов (а не лингвистические)
 - Простые признаки текстов такие, как слова в отличие от синтаксического разбора и учета предложений
 - Лингвистические признаки могут быть частью статистической модели

Оценка качества поиска (evaluation)

- Экспериментальные процедуры и меры для сравнения результатов работы систем с ожиданиями пользователей
- Метода оценки качества поиска сейчас используются во многих областях
- Типично используются тестовые коллекции документов, запросов, и оценки релевантности
- *Полнота и точность – простые примеры оценки качества*

Пользователи и информационная потребность

- Ключевые слова – это слишком бедное описание действительных информационных потребностей
- Взаимодействие и контекст – важны для понимания потребности пользователя
- Методы уточнения запроса: расширение запроса, предложение запроса, *relevance feedback*

Информационный поиск: основные проблемы

- Построение представления содержания документа
- Построение описания потребности пользователя
- Сравнение представления содержания документа и представления потребности пользователя
- Оценка эффективности информационного поиска
- Интернет vs. Интранет

Конкретные приложения

Интернет-поиск

Яндекс

автоматическая обработка текстов — 9 млн ответов

Найти

Почта



Поиск



Картинки



Видео



Карты



Маркет



W Категория: [Автоматическая обработка текстов](#) — Википедия

ru.wikipedia.org > Категория: [Автоматическая обработка текстов](#) ▼

Подкатегории. В этой категории отображается 3 подкатегории из имеющихся 3. О. ► **Обработка естественного языка** (32: 5 кат., 27 с.) П. [×] **Автоматизированный перевод** (14: 14 с.) [×] **Преобразователи текста** (2: 2 с.) Страницы в категории «**Автоматическ...**

Разместить объ
«автоматическа
показа в месяц

Все объявления

АОТ :: Главная

aot.ru ▼

АОТ. Автоматическая Обработка Текста. ... © 2003 АОТ. Все права защищены.

А Автоматическая обработка текстов — Школа анализа данных

shad.yandex.ru > Школа анализа данных > [text.xml](#) ▼

Лекция 1. Введение в автоматическую обработку текстов. Стандарт Unicode. Токенизация и нормализация текста.

А Автоматическая обработка текста

tapemark.narod.ru > [les/014a.html](#) ▼

Автоматическая обработка текста —. преобразование текста на искусственном или естественном языке с помощью ЭВМ.

А Автоматическая обработка текстов на естественном языке...

window.edu.ru > Учебное пособие > [.../miem_lingvistika.pdf](#) ▼

Министерство образования и науки Российской Федерации Московский государственный институт электроники и математики **АВТОМАТИЧЕСКАЯ ОБРАБОТКА ТЕКСТОВ НА...**



5.7 МБ

Посмотреть

Загрузить

А tpl-it - Автоматическая обработка текста


tpl-it.wikispaces.com > [Автоматическая обработка...](#) ▼

Определение из Энциклопедии русского языка: **Автоматическая обработка текста** - преобразование текста на ... В зависимости от целей различают несколько видов АОТ.

Подзадачи информационного поиска

- Исправление опечаток в запросе
- Расширение запроса
- Диверсификация выдачи поисковой системы
- Гипертекст и ссылки
- Логи запросов
- Клики
- Определение дубликатов
- Поисковый спам и др.

Диверсификация выдачи поисковой системы



Нашлось
27 млн ответов

Поиск

Почта

Карты

Маркет

Новости

Словари

Блоги

Видео

Картинки

ещё

✕

Найти

☐ в найденном ☐ в Москве

расширенный поиск

Мои находки

Настройка

Регион: Москва

По

Результаты

все

в рунете

в мировом интернете

1

Новые смартфоны **Nokia** - ...смартфоны на базе Windows Phone - **Nokia...**

Телефоны

Аксессуары

Russia

Lumia Популярные приложения

Приложения

Магазины

Официальная страница **Nokia** Россия. Посетите сайт, чтобы открыть весь мир новых смартфонов **Nokia** на базе Windows Phone **Nokia**!

Facebook

YouTube

Twitter

ВКонтакте

ежедн. 9:00-23:00 +7 (495) 967-91-00

Москва Домодедовская Каширское ш., 61, корп.2

nokia.com > Russia

копия

ещё

2

shop.nokia.ru/

shop.nokia.ru

3

All **Nokia** - Клуб любителей телефонов **Nokia** / Скачать бесплатно для Нокиа...

Телефоны

Прошивки

Инструкции

Новости

Программы

Темы

Обзоры телефонов и смартфонов, отзывы пользователей. Возможность скачать игры, музыку, темы, программы. Статьи и инструкции.

allnokia.ru

4

Nokia — Википедия

Nokia (произносится Но́киа) — финская транснациональная компания, производитель мобильных телефонов, смартфонов, а также телекоммуникационного оборудования для мобильных, фиксированных, широкополосных и IP-сетей.

ru.wikipedia.org > Википедия > Nokia

5

Телефоны **Nokia**. Каталог мобильных телефонов **Nokia** (Нокиа). Новинки...

Новинки

Nokia Lumia 720

Nokia Lumia 520

Яндекс.Директ

Смартфоны **Nokia** в Связном!

Покупайте смартфоны **Nokia** в интернет-магазине Связной! Кредит от 0%!

svyaznoy.ru

Все объявления

Разместить объявление по запросу «nokia» — 3 412 175 показов в месяц

Видео «nokia»

Lumia 920 vs iPhone 5

10:16

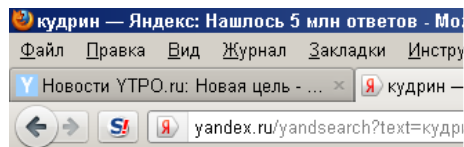
Nokia Lumia 920 против iPhone 5. Сравнение AppleInsider.ru Nokia Lumia..

Все видеоролики

Вертикальный поиск

- Вертикальный поиск - тематический поиск в Интернет: книги, недвижимость, новости, карты и др.
- Современные поисковые системы – включают элементы вертикального поиска: некоторые запросы направляют на соответствующие вертикали
- Яндекс: поисковые колдунщики
- Вертикали Яндекса: погода, конвертер валют, википедия, новости, маркет, картинки, видео, музыка, вакансии и др.
- <http://help.yandex.ru/search/?id=1111313>

Вертикальный поиск по НОВОСТЯМ



Яндекс
Нашлось
5 млн ответов

[Поиск](#) [Почта](#) [Карты](#) [Маркет](#) [Новости](#) [Словари](#) [Блоги](#) [Видео](#) [Картинки](#) [ещё](#)

кудрин

Найти

☐ в найденном ☐ в Москве

[расширенный поиск](#)

[Мои находки](#)

[Войти](#)

[Настройка](#)

[Регион: Москва](#)

[Все объявления](#)

[Уход Алексея Кудрина](#)

"Единая Россия" сделала Кудрина мишенью для критики. Читайте на сайте www.firstnews.ru



[Чубайс: «Последствия отставки Кудрина могут оказаться драматическими»](#)



Гендиректор ОАО "Российская корпорация нанотехнологий" Анатолий Чубайс заявил, что отставка вице-премьера, министра финансов России Алексея Кудрина может иметь драматические последствия. Решение об отставке главы Минфина, "какими бы причинами оно не было вызвано, создает серьезные риски для страны", отметил А. Чубайс в своем блоге.

[Эхо Москвы](#) 14:18 [Финмаркет](#) 12:03 [Взгляд.ру](#) 10:36

[Все сообщения](#) 26

[Путин утвердил новое распределение обязанностей в правительстве](#) 295 сообщений

[Ясин: Отставка Кудрина грозит ростом расходов и дефицита бюджета](#) 2050

сообщений

news.yandex.ru 4 часа назад



[Кудрин, Алексей Леонидович — Википедия](#)

[Биография](#) [Экономическое мировоззрение](#) [Библиография](#)

Алексе́й Леони́дович Ку́дрин (12 октября 1960, Добеле, Латвийская ССР) — российский государственный деятель, министр финансов Российской Федерации (с мая 2000)...

ru.wikipedia.org > [Кудрин](#) [копия](#) [ещё](#)

[Яндекс.Директ](#)

[Кудрин ушёл в отставку!](#)

Причины ухода министра финансов. Последствия для экономики. Новости часа www.bfm.ru

[Министр финансов Кудрин](#)

отправлен в отставку! Что за этим последует? Подробности: www.zagolovki.ru

[Почему Кудрин не хочет работать в](#)

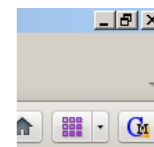
команде Медведева? Подробности читайте на "Голосе Америки" www.voanews.com

[Отставка Кудрина](#)

Чем грозит отставка госбюджету России? Узнайте подробнее на: investcafe.ru

[Разместить объявление по запросу «кудрин»](#) — 15 519 запросов в месяц

[Видео «кудрин»](#)



Расширение запросов

- Слова запроса могут не соответствовать словам документа
- Как искать?
 - Подсказки
 - Расширение запроса, т.е. поиск не только по словам запроса, но и по близким по смыслу к словам запроса
 - Тезаурусы – специальные лингвистические ресурсы
 - Автоматически полученные взаимосвязи

Нашлось
554 тыс.
ответов

☐ в найденном ☐ в Троицке[расширенный поиск](#)

[Как очистить кожу лица в домашних условиях](#)

Очищать кожу можно дважды в день после сна утром и перед сном вечером. ... Если вы не будете **очищать кожу**, это приведет к черным точкам и прыщам.

[www.AllWomens.ru/5349-kak-ochistit-kozhu-lica-v-...](#) [копия](#) [ещё](#)

[Очистка кожи Восток сервис. Крем-паста НАША ФОРМУЛА для очистки...](#)

Рекомендуется для **очистки кожи** рук и тела от легких производственных загрязнений. Отличительные характеристики: жидкое **очищающее** средство мягкого действия для удаления с ...

[www.vostok.ru/catalogue/5679/](#) [копия](#) [ещё](#)

[Средства для очистки кожи - Средства от прыщей - Happy-face.ru](#)

Смотря на обилие разного рода косметических средств для **очистки кожи**, которых полно на прилавках магазинов, становится немного жалко женщин, которые жили в более старые времена.

[happy-face.ru/sredstva-dlya-ochistki-kozhi/](#) [копия](#) [ещё](#)

[Народные средства очистки кожи лица Забота о здоровье и красоте](#)

Перечисленные ниже народные средства **очистки кожи** предназначены прежде всего для сухой и нормальной **кожи**, а также для чувствительной ее разновидности.

[www.soul-body.ru/index.php/?narodnye...ochistki...html](#) [копия](#) [ещё](#)

[Способы очистки кожи. Основная очистка лица](#)

Недостаток жирной **кожи** обычно заключается в чрезмерной щелочности, поэтому нужно постоянно использовать для **очистки** кислые продукты, хорошо **очищающие** кожу и уменьшающие ее раздражение.

[Разместить объявление по запросу «очистки-кожуры»](#)

[«очистки-кожуры» в картинках](#)



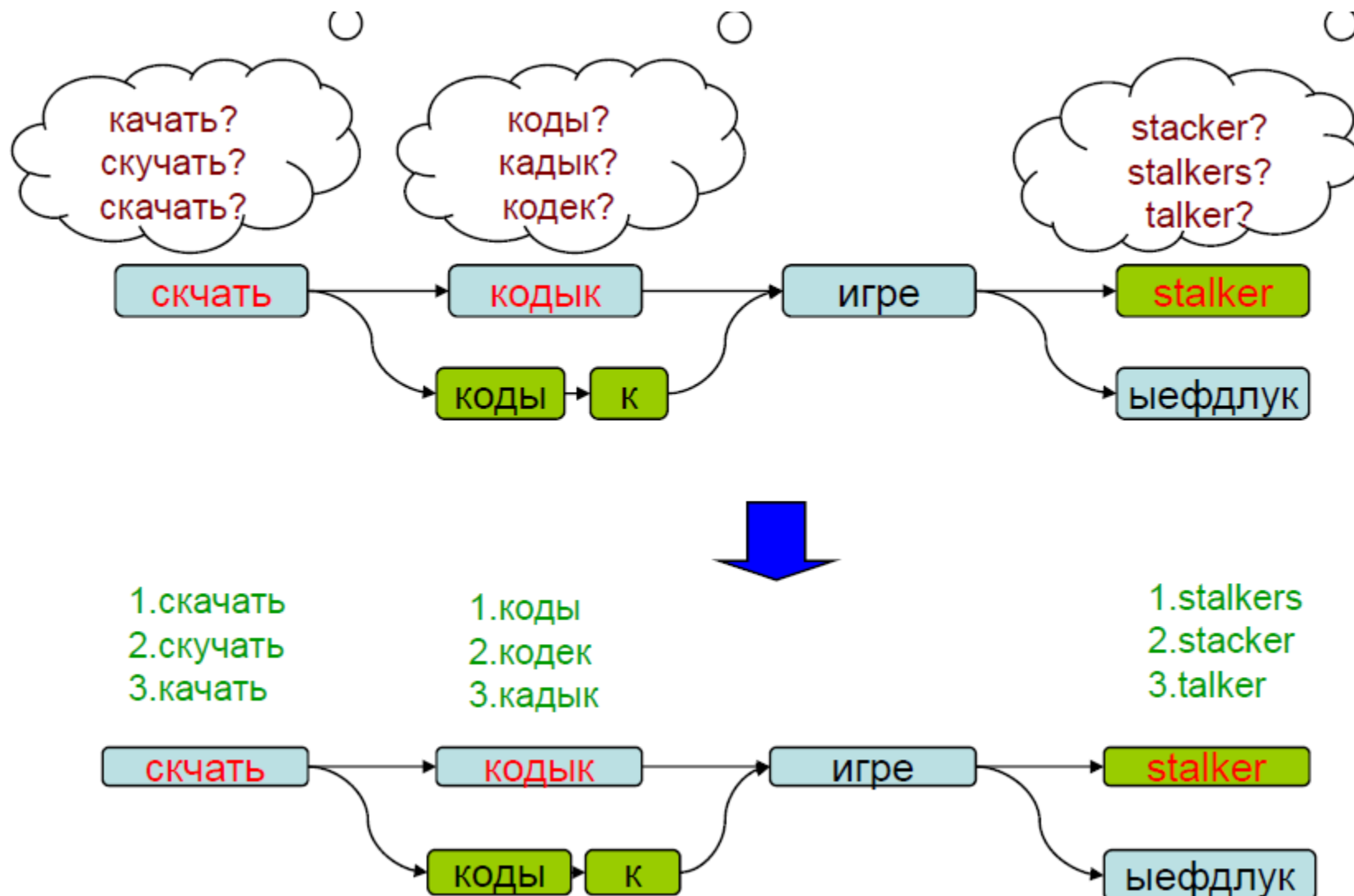
[Все картинки](#)

[Видео «очистки-кожуры»](#)



[Очистка кожи лица](#)
[Все видеоролики](#)

Исправление опечаток в запросе



Вопросно-ответный поиск

Ответы на вопросы –

сравнительно новая задача, актуальная

(но и забытое старое направление, 70 гг.)

- Нужен не документ или сниппет, а ответ на конкретный вопрос ,
например: *Кто придумал вилку?*
- Примерная стратегия построения ответа:
 - определение типа вопроса
 - построение запроса к интернет-поисковику
 - извлечение из найденных документов нужной информации
 - построение фразы ответа



где родился обاما

Все

Новости

Видео

Картинки

Карты

Ещё ▾

Инстру

Результатов: примерно 2 240 000 (0,69 сек.)

Барак Обама / Место рождения

Гонолулу, Гавайи, США

Оставить отзыв

Обама, Барак — Википедия

https://ru.wikipedia.org/wiki/Обама,_Барак ▾

Бара́к Хуссе́йн **Оба́ма** II (англ. Barack Hussein **Obama** II, произносится [bəˈrɑːk ... **Родился** в городе Гонолулу, штат Гавайи. Его родители познакомились в 1960 году во время учёбы в Гавайском университете в Маноа. Вместе с ...

[Методизм](#) · [Данхэм, Энн](#) · [Обама, Барак Хусейн](#) · [Мишель](#)

[Первая девушка Обамы выдала тайну президента США //](#)

Яндекс

где родился обاما



Поиск

Картинки

Видео

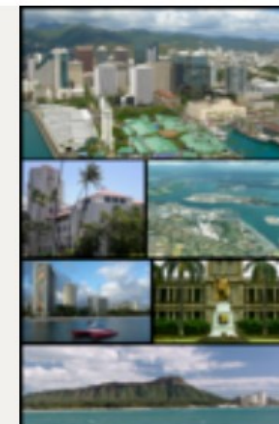
Карты

Маркет

Ещё

где родился обاما

Гонолулу



 **Обама, Барак** — Википедия

ru.wikipedia.org > **Обама**, Барак ▼

Детство, образование, начало карьеры. **Родился** в городе Гонолулу, штат Гавайи. ... В 1988 году **Обама** поступил в школу права Гарвардского университета, **где** в 1990 году...



Барак Обама. Биография президента. Barack Obama.

states-of-america.ru > Личности США > Барак **Обама** ▼

Когда Барак **Обама** только **родился**, отец уехал в Гарвард, чтобы ... Два года спустя, он переехал в Кению, **где** ему предложили работу в аппарате правительства.



Барак Обама биография

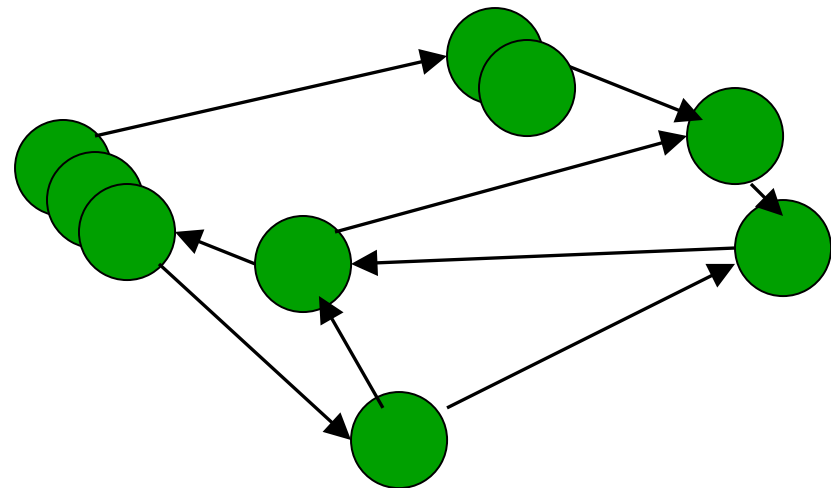
to-name.ru > Ксения > Барак **Обама** ▼

Настоящее время: вопросно-ответные системы и диалоговые системы

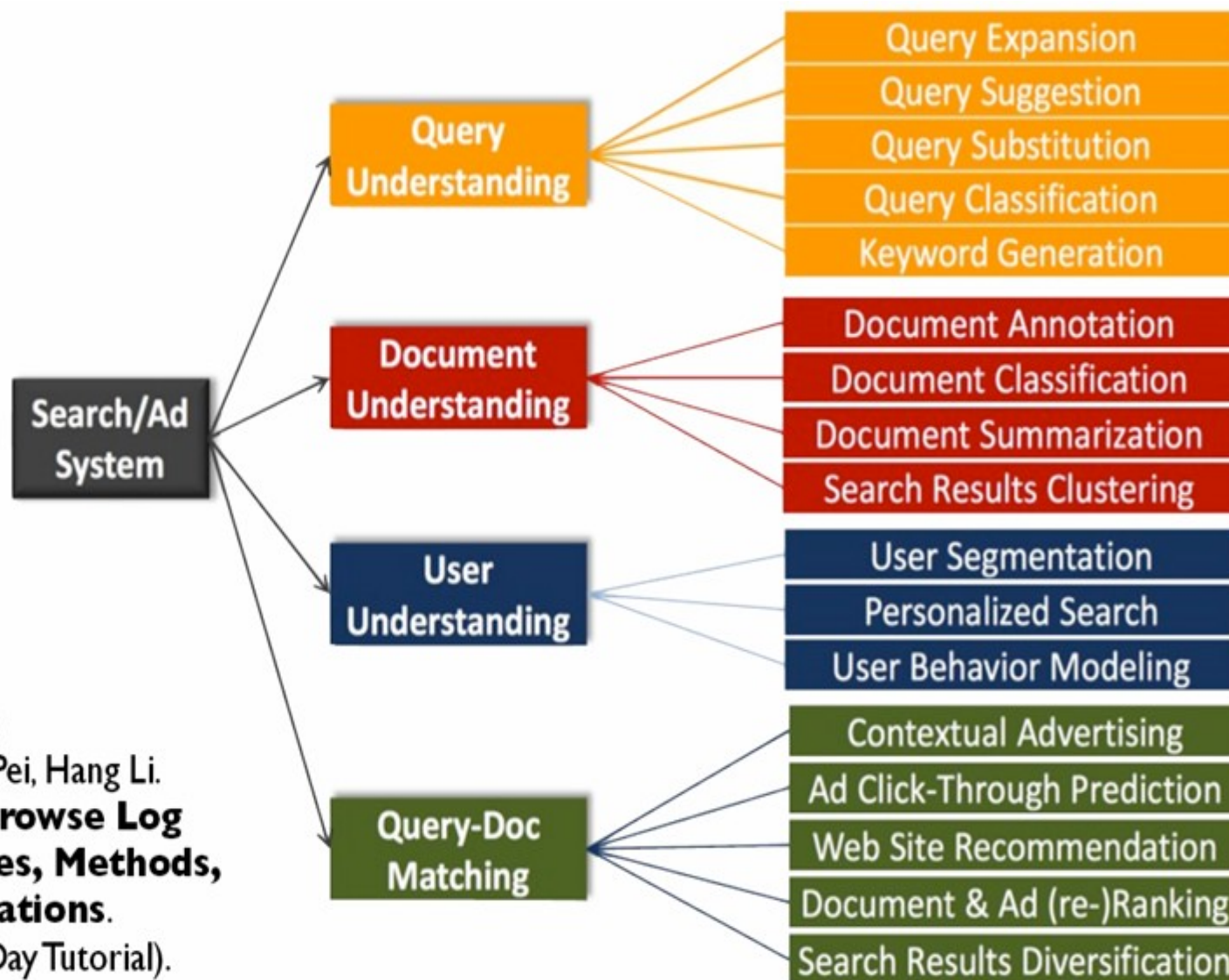
- Диалоговые системы
 - Вопросно-ответная система, которая «помнит» прошлое
 - Разрабатываются много лет, но сейчас новые возможности: много данных, роль чатов (письменных коротких сообщений) возросла, рост
- Ответы на вопросы пользователя в колл-центрах
 - Включает поиск в существующих вопросах-ответах
- Чат-боты
 - поддерживают беседу на разные темы
 - Часто должны находить ответы среди имеющихся, т.е. **применять информационный поиск**

Учет ссылок между страницами

- Что есть кроме содержания документов
 - Гиперссылки между документами
- Вопросы
 - Могут ли ссылки продемонстрировать авторитетность страниц? Полезны ли они для ранжирования?
- Применение
 - Интернет
 - Email
 - Социальные сети

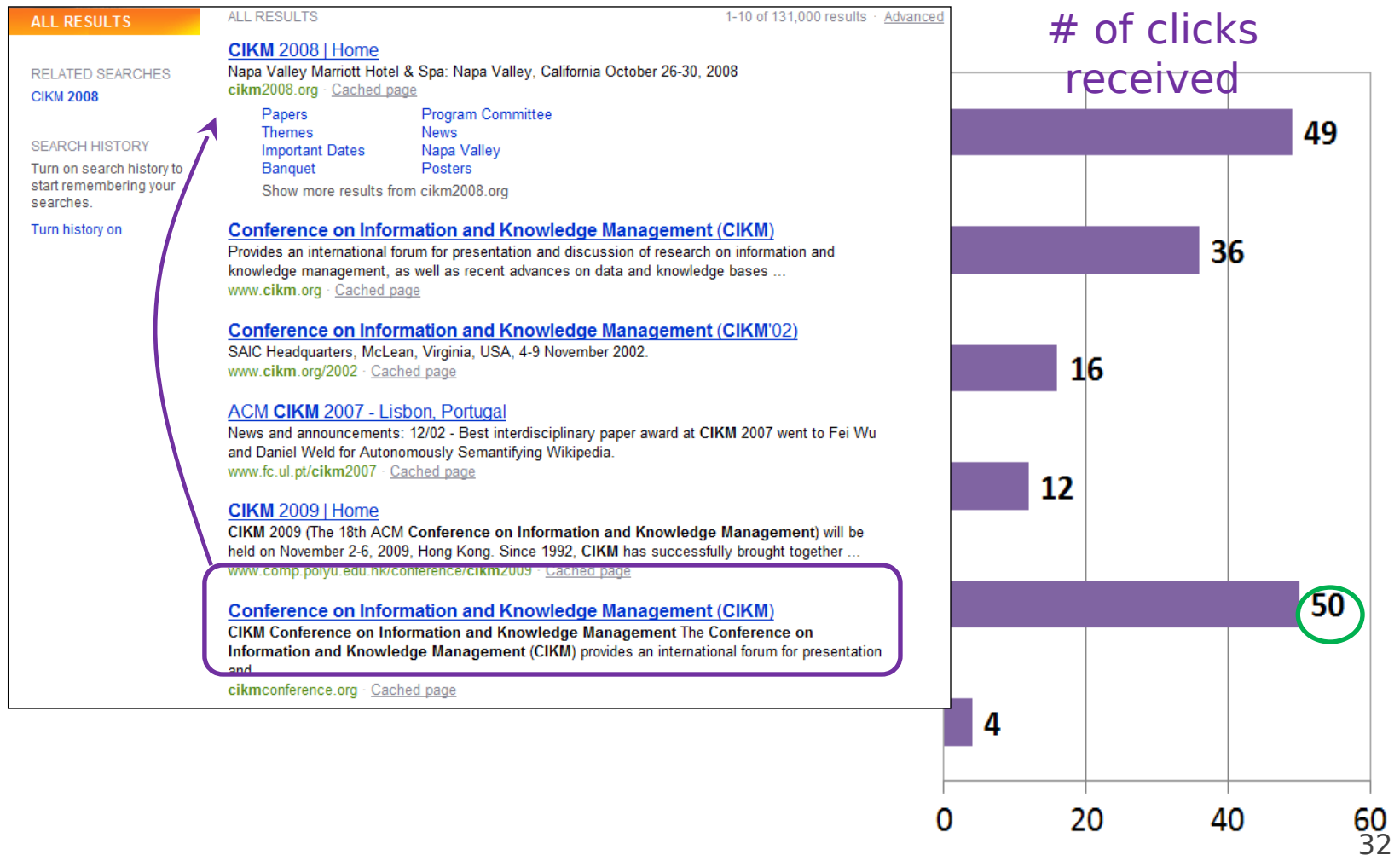


Логи пользователей и их использование

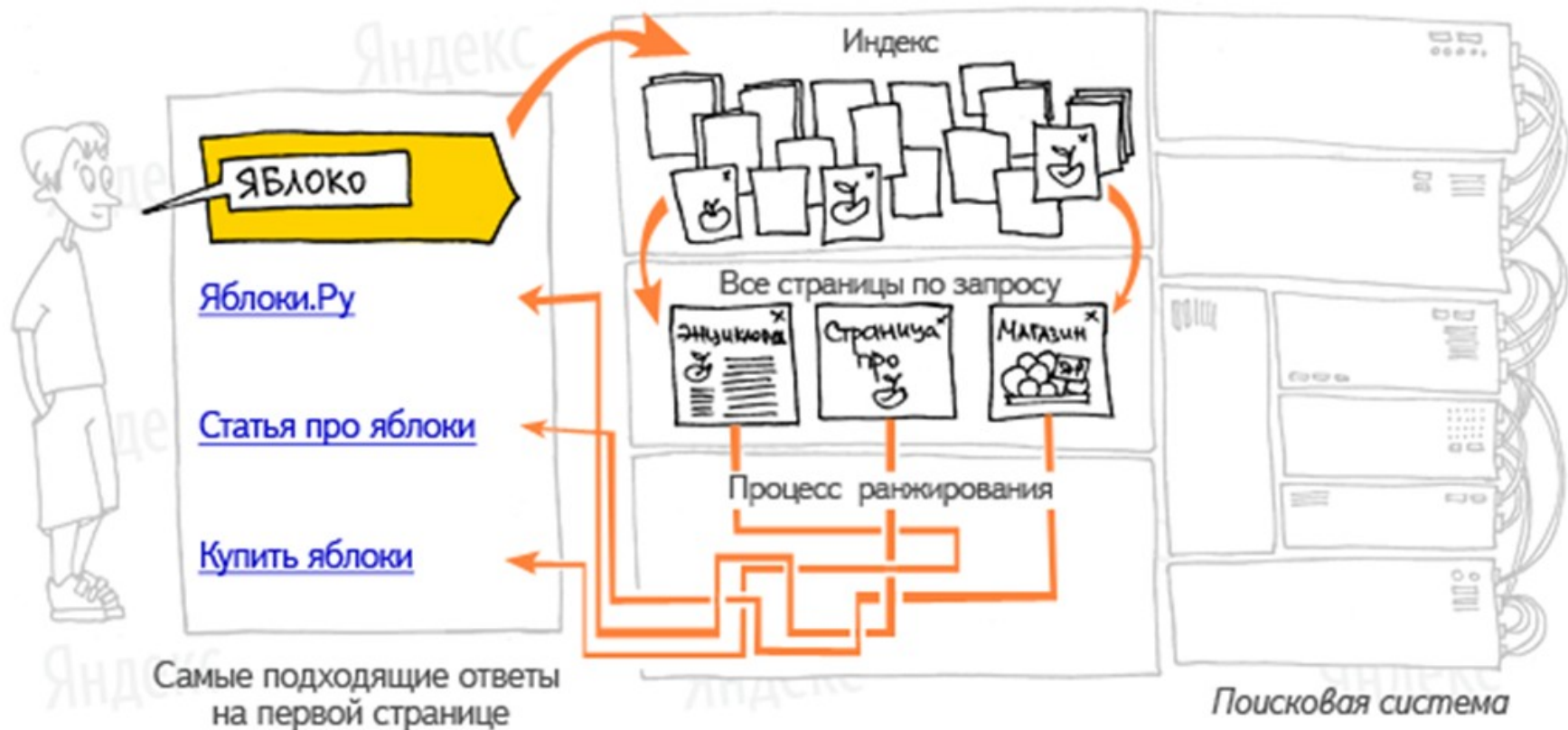


From:
Daxin Jiang, Jian Pei, Hang Li.
Web Search/Browse Log Mining: Challenges, Methods, and Applications.
WWW'10 (Full-Day Tutorial).

Как улучшить выдачу используя клики?

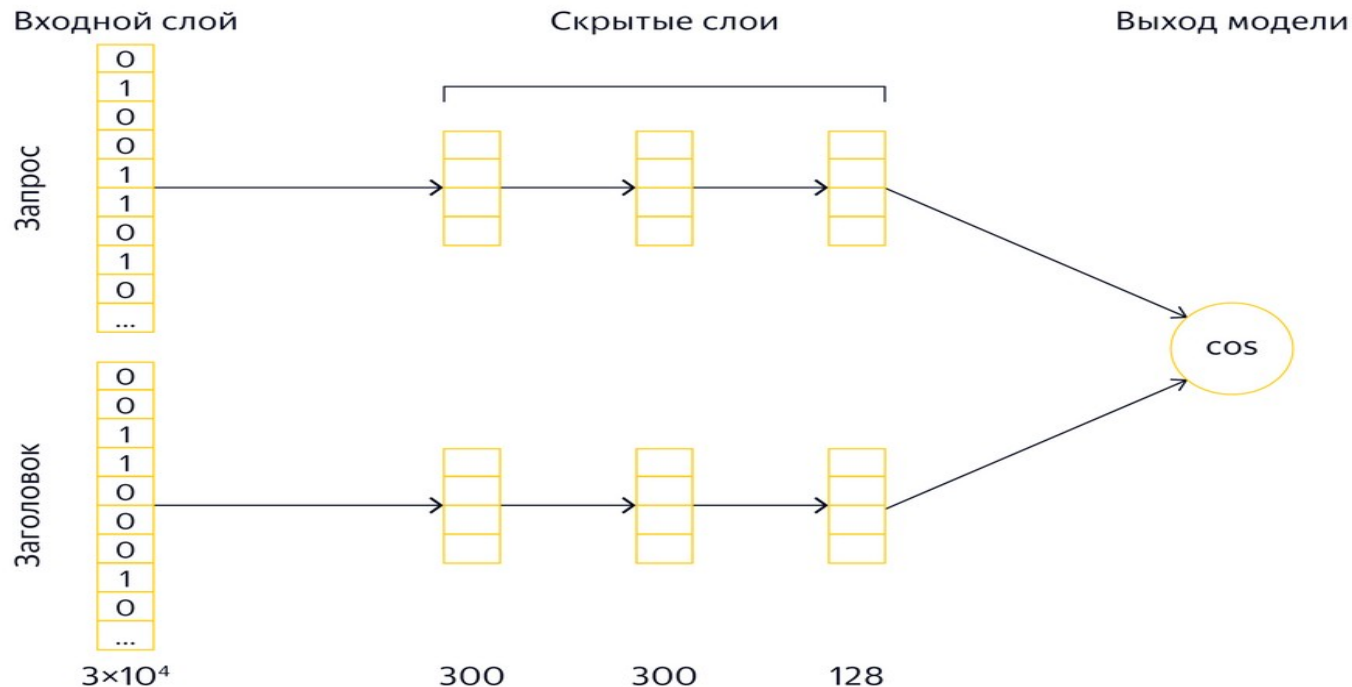


Комбинирование большого числа признаков (запросов, пользователей и документов): MatrixNet от Яндекса



Новые алгоритмы Яндекса:

- На основе нейронных сетей
 - Палех (ноябрь 2016) (сравнение запроса с заголовком)
 - Королев (август 2017) (сравнение запроса с полным текстом)
- Deep structured semantic model (DSSM)
 - Текст в виде 300 чисел



Другие задачи

- Автоматическая классификация (рубрикация) документов
- Автоматическая кластеризация документов
- Автоматическое аннотирование

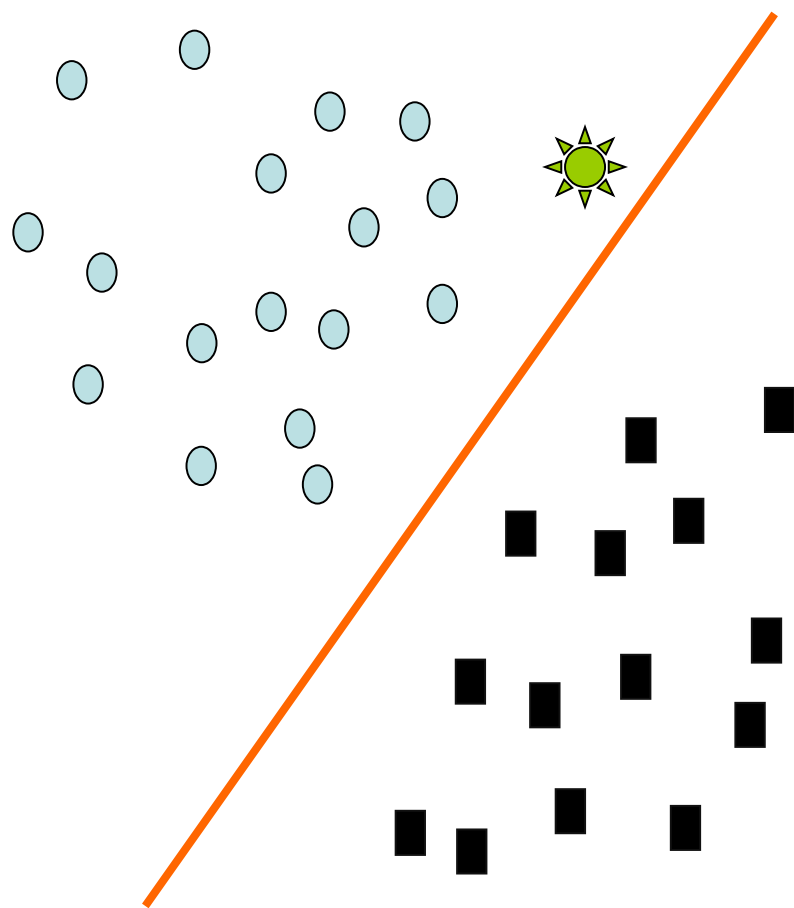
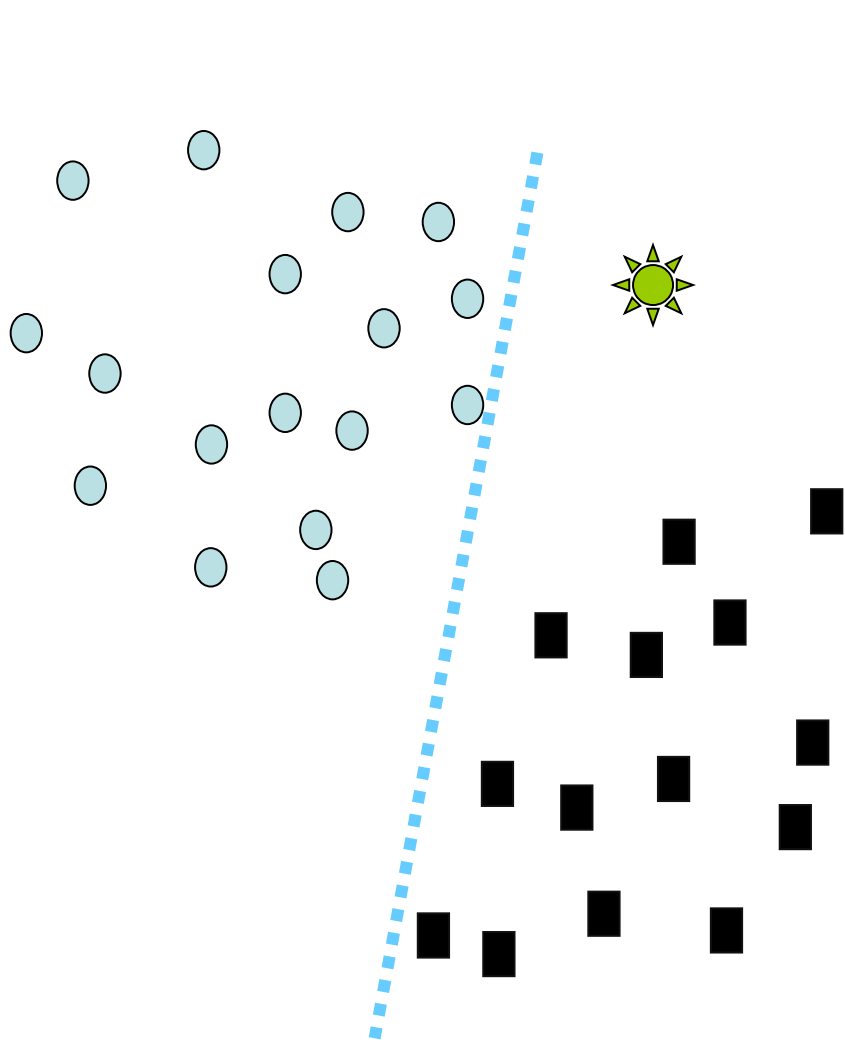
Классификация (рубрикация) документов

- Классификация/рубрикация информации – отнесение порции информации к одной или нескольким категориям из конечного множества рубрик
- Применение:
 - Навигация по коллекции документов
 - Поиск информации
 - Замена сложного запроса
 - Иерархическое упорядочение знаний предметной области
 - Анализ распределения документов по тематике
 - Фильтрация потока текстов:
 - Тематический сбор новостей
 - Персонализированная фильтрация потока текстов
 - Фильтрация спама
 - Тематический сбор информации из интернет

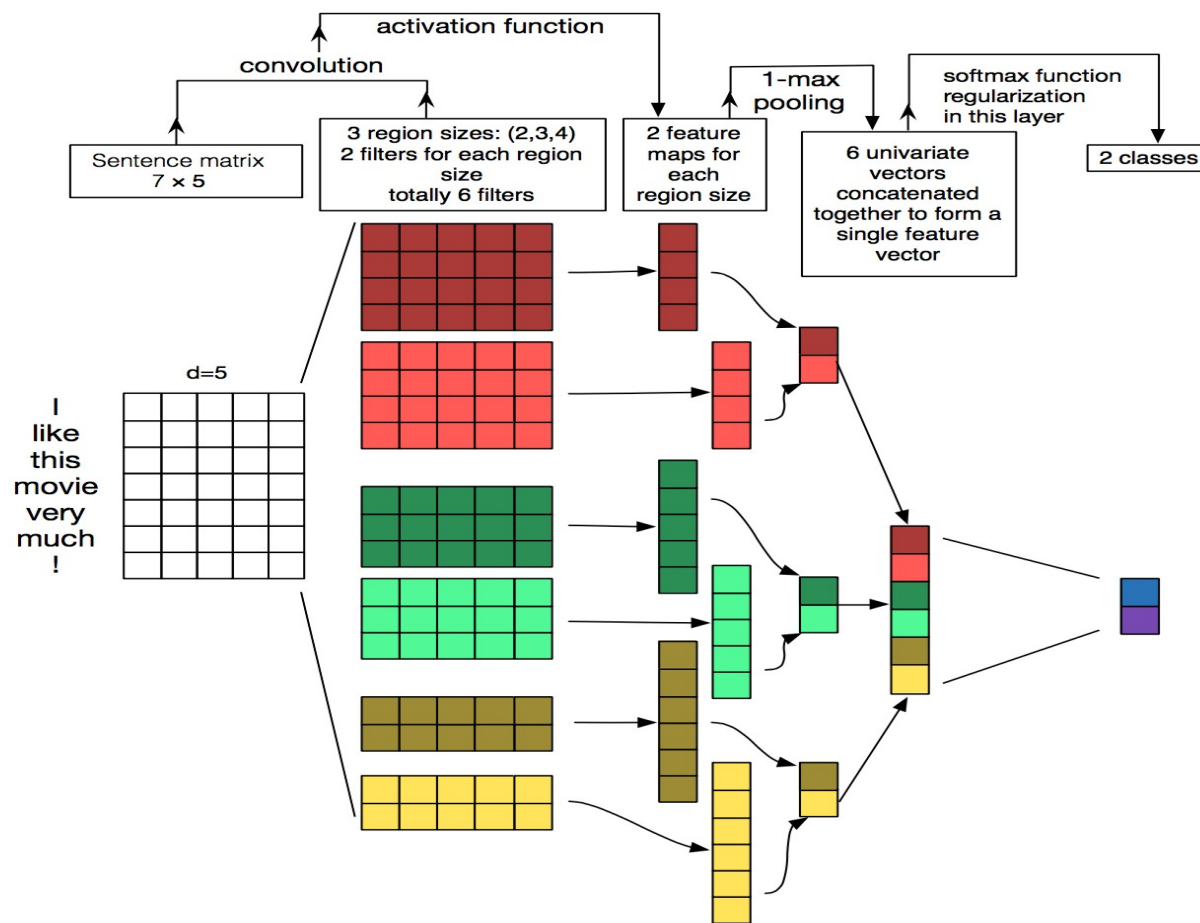
Каталог Яндекс – Фасетная классификация

- **Тематическая**
 - Иерархический классификатор, имеет порядка 600 значений и описывает предметную область интернет-ресурса
- **Регион**
 - 230 географических областей. Определяется географическим расположением представляемого объекта, сферой управления и влияния, потенциальной аудиторией информации или информационным содержанием ресурса
- **Жанр**
 - художественная литература; научно-техническая литература; научно-популярная литература; нормативные документы; советы; публицистика
- **Источник информации**
 - Официальный, СМИ, Неформальный, Персональный Анонимный
- **Адресат информации**
 - Партнеры, Инвесторы, Потребители, Коллеги
- **Сектор экономики**
 - Государственный, Коммерческий, Некоммерческий

Положительные и отрицательные примеры: как лучше отделить



Нейронные сети в задачах классификации текстов



Классическая нейронная сеть для классификации текстов по тональности (2014)

Автоматическая кластеризация текстов

- Имеется текстовая коллекция
- Нужно разбить коллекцию на классы близких документов
- Могут быть созданы иерархические классы
- Сейчас: одно из важных средств для визуализации большой выдачи документов при поиске
- Для визуализации важно: хорошее название кластера
- Примеры:
 - Новостные агрегаторы (Яндекс.Новости, Рамблер.Новости, Google.News, Новотека)
 - Кластеризация результатов поиска (Clusty, Нигма)

Антидопинговые агентства 17 стран призвали к реформе WADA

Лидеры 17 национальных антидопинговых организаций встретились на саммите в Копенгагене и предложили реформировать Всемирное антидопинговое агентство (WADA), сообщает Би-би-си. В частности, руководителям WADA предлагается запретить занимать важные позиции в других спортивных организациях. РБК 00:32

С чего всё началось

Anti-doping leaders demand Wada overhaul with clean sport "at crossroads" - BBC Sport

An overhaul of the World Anti-Doping Agency (Wada) is proposed in a bid to restore trust in international sport. www.bbc.com

ПОДРОБНЕЕ О СОБЫТИИ

S Sports.ru
13:40

Антидопинговые агентства 17 стран выступили за реформирование WADA

IN Народные новости
11:41

Перемены в WADA ударят по зачинщикам реформ

С Коммерсантъ-Online
11:24

WADA нуждается в реформах

С Sntat.ru
09:18

Главы антидопинговых агентств 17 стран заявили о необходимости реформы WADA

L Lenta.ru
02:32

Главы антидопинговых агентств 17 государств призвали реформировать WADA

В Вести.Ru
01:09

В Копенгагене обсудили реформирование WADA

РБК РБК
00:32



Фото 75 Видео 7

ОБЩЕСТВО

В аэропорту Франкфурта-на-Майне задержана подозрительная пассажирка

Учительнице из Татарстана предъявили обвинение за секс с ученицей

Более 15 пассажиров United Airlines пострадали во время турбулентности

Названы города России с худшей экологической обстановкой

Парламент поддержит поправки о приоритете зарплат перед налогами

ДАНИЯ

Датские СМИ испугались нового российского супероружия

Власти Дании намерены отказывать просителям убежища прямо на границе

В столице Дании произошли новые поджоги автомобилей

Дания собирается усилить свою роль в Арктике

Правительство Дании предложило разворачивать просителей убежища на границах страны

ГЛАВНЫЕ НОВОСТИ



Ярославский

Квартиры
от 2,3 млн руб.



www.pik.ru


Застройщик ООО «Загородная усадьба».
Проектные декларации размещены на сайте pik.ru

Автоматическое аннотирование

- Чаще всего наиболее содержательный фрагмент документа
 - Индикативная аннотация – пересказ основного содержания текста
 - Контекстно-зависимая аннотация
 - сниппеты в поисковых системах
 - Аннотация многих документов
 - Аннотация новостных кластеров

Контекстные аннотации: сниппеты

Поиск [Почта](#) [Карты](#) [Маркет](#) [Новости](#) [Словари](#) [Блоги](#) [Видео](#) [Картинки](#) [ещё](#)



Нашлось
29 тыс. ответов

☐ в найденном ☐ в Москве


Найти

расширенный поиск


[Мои нах](#)
[Настрой](#)
[Регион:](#)

[Разместить объявле](#)
[«автоматическое...»](#)
месяц


1

 [Автоматическое реферирование статей на русском языке / Хабрахабр](#)
5 мая 2011 Тема **автоматического реферирования/аннотирования** текста была поднята давно и было придумано множество способов ее реализации.
[habrahabr.ru](#) > [Писалось](#)


2

 [Автоматическое реферирование и аннотирование](#)
«Либретто» (разработчик — компания «МедиаЛингва»), обеспечивающую **автоматическое реферирование** и аннотирование русских и английских текстов (система встраивается в Word)
[do.gendocs.ru](#) > [docs/index-13506.html](#)


3

 [Автоматическое реферирование и аннотирование](#)
Автоматическое реферирование и аннотирование — одно из направлений компьютерной обработки естественно-языковых текстов.
[do.gendocs.ru](#) > [docs/index-208893.html](#)


4

 [автоматическое реферирование](#)
Большой англо-русский и русско-английский словарь. automated abstracting — Лингвистика: **автоматическое аннотирование, автоматическое реферирование** ...
[dic.academic.ru](#) > [dic.nsf/eng_rus...автоматическое](#)

5

 [Системы автоматического реферирования](#)
Системы **автоматического реферирования**. Искусство **реферирования**, или составления аннотаций, или кратких изложений материала, иными словами...
[rudocs.exdat.com](#) > [docs/index-34660.html](#)


6

 [Автоматическое реферирование](#)
Автоматическое реферирование (Automatic Text Summarization) - это составление коротких изложений материалов, аннотаций или дайджестов, т.е...
[bourabai.kz](#) > [dbt/internetica/autorefer.htm](#)

7

[Системы автоматического реферирования](#)
Системы **автоматического реферирования**. Удо Хан, Индерджуиет Мани, "Открытые Системы" #12/2000. В статье рассматриваются инструменты и методы **реферирования**...

Аннотирование новостного кластера

**Яндекс**
НОВОСТИ


Поиск Почта Карты Маркет **Новости** Словари Блоги Видео Картинки ещё

☐ только в этом сюжете [расширенный поиск](#)

[Войти](#) [Помощь](#)

Главные новости Мои новости Политика В мире Общество **Экономика** Спорт Происшествия Культура Наука Hi-Tech Интернет Авто


На помощь Дальнему Востоку решено выделить 12 млрд рублей

Интерфакс  **Дальнему Востоку выделено 12 млрд рублей** 15:29


На заседании правительства России, состоявшемся в среду, принято решение выделить 12 млрд рублей на оказание материальной помощи, восстановление инфраструктуры и развитие пунктов временного размещения для пострадавших от наводнения на Дальнем Востоке.

Взгляд.ру **В** **На помощь Дальнему Востоку решено выделить 12 млрд рублей** 14:02


Бывший глава Минприроды, помощник президента **Юрий Трутнев** 31 августа был назначен вице-премьером и одновременно полпредом в Дальневосточному федеральному округу.

РИА Новости  **Федеральный центр выделяет 12 млрд рублей на помощь Дальнему Востоку** 13:41


"Правительство приняло решение о выделении 12 миллиардов рублей на оказание материальной помощи пострадавшим, проведение аварийных работ и временное размещение людей", — сказал **Трутнев** журналистам по итогам заседания кабинета министров.



Карта



Все видео




Все фото

Ещё по теме

Трутнев не исключает увеличения федеральной помощи Дальнему Востоку 14:25

Зейская и Бурейская ГЭС задержали 65% паводка в Амурской области 12:40

В Красноярске создадут штаб

 **GE Money Bank**

13,5%

Для женщин

Извлечение информации из текстов

- Именованные (конкретные) сущности

NE - Named Entities

- персоны, компании, адреса, даты
- упоминания генов и белков и пр.

- Отношения выделенных сущностей:

- Место работы, должность
- Взаимодействие белков

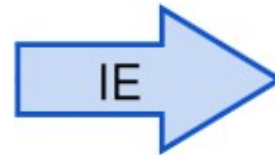
- Связанные с ними события и факты

Events

слияние/поглощение компаний...

приобретение контрольного пакета акций

Извлечение отношений: из неструктурированной формы в структурированную



Subject	Relation	Object
p53	is_a	protein
Bax	is_a	protein
p53	has_function	apoptosis
Bax	has_function	induction
apoptosis	involved_in	cell_death
Bax	is_in	mitochondrial outer membrane
Bax	is_in	cytoplasm
apoptosis	related_to	caspase activation
...

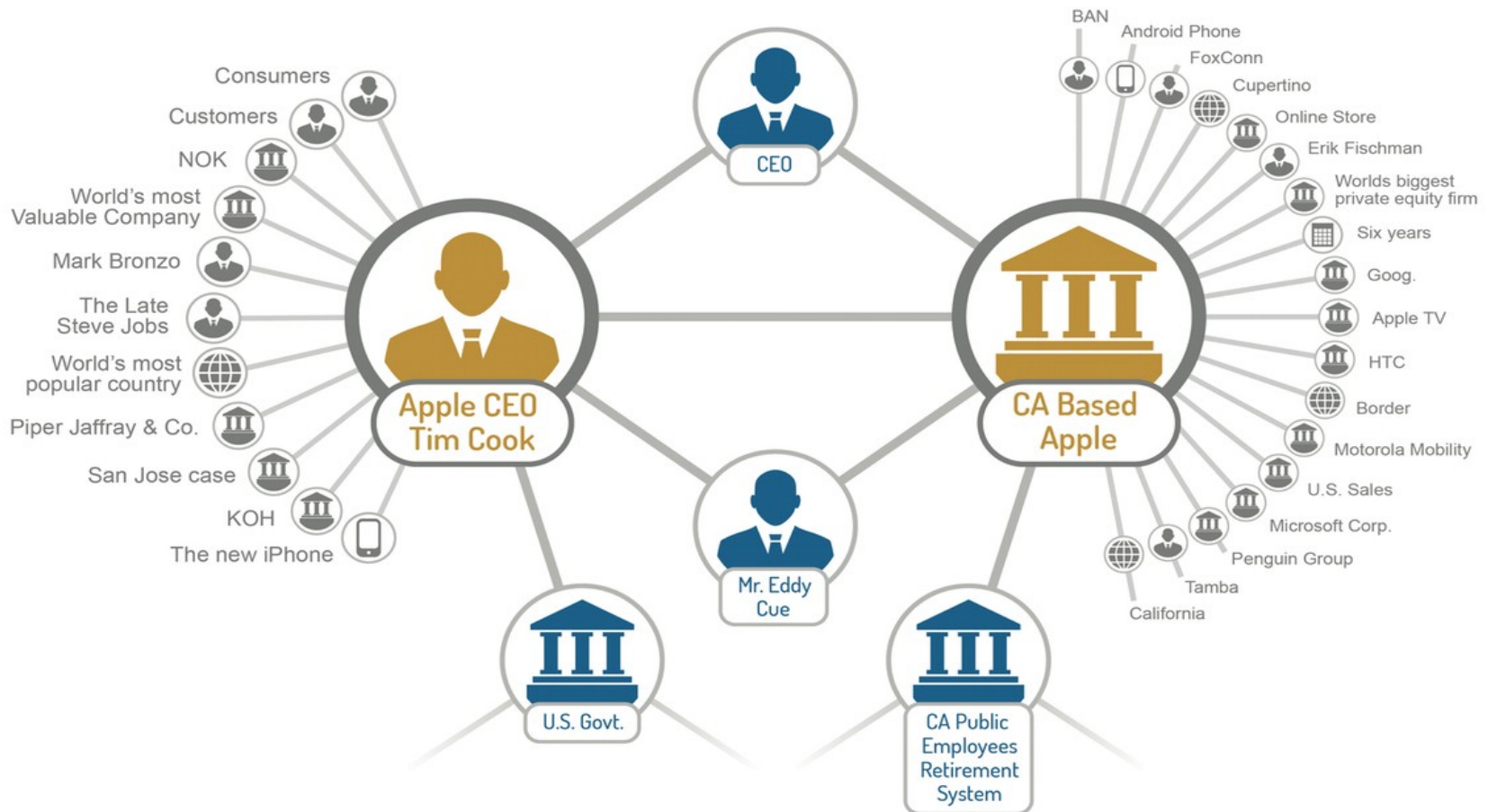
Типичные отношения извлекаемые из новостей

Relations		Examples	Types
Affiliations	Personal	<i>married to, mother of</i>	PER → PER
	Organizational	<i>spokesman for, president of</i>	PER → ORG
	Artifactual	<i>owns, invented, produces</i>	(PER ORG) → ART
Geospatial	Proximity	<i>near, on outskirts</i>	LOC → LOC
	Directional	<i>southeast of</i>	LOC → LOC
Part-Of	Organizational	<i>a unit of, parent of</i>	ORG → ORG
	Political	<i>annexed, acquired</i>	GPE → GPE

Графы знаний в современных поисковых системах

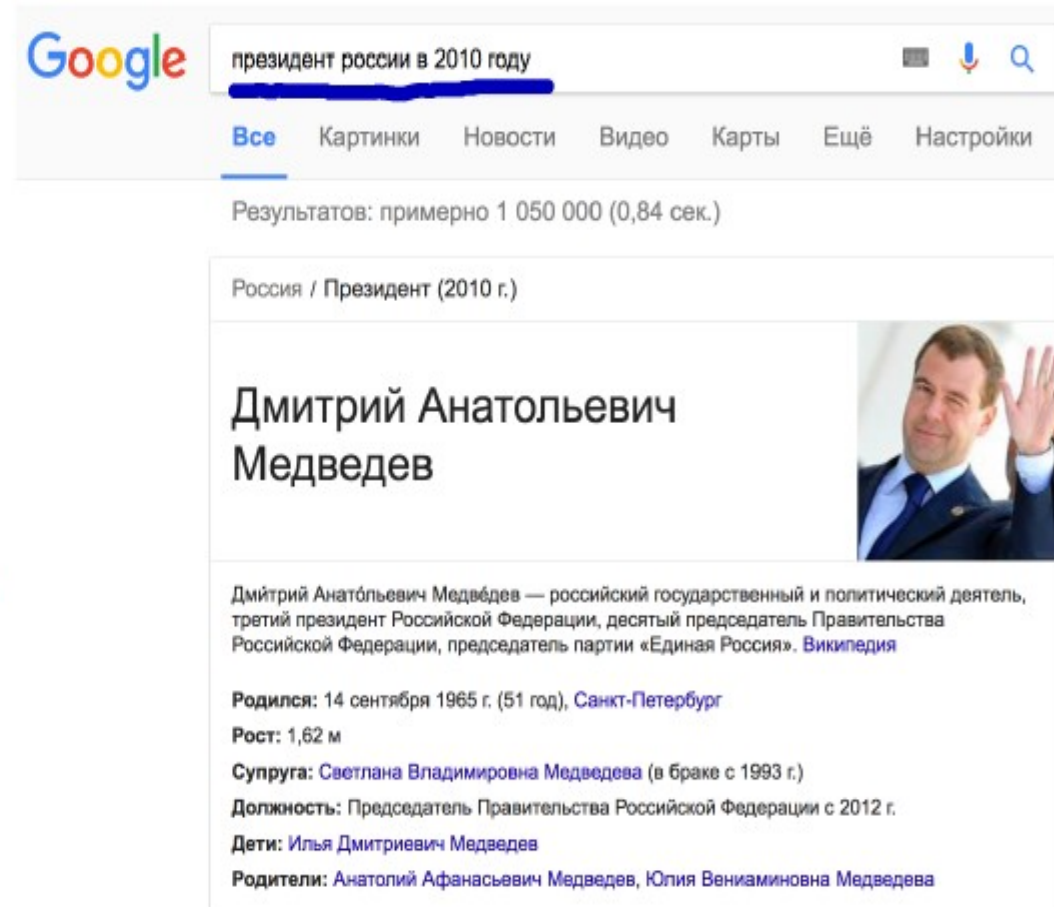
THE KNOWLEDGE GRAPH

LINKS TOGETHER BILLIONS OF ENTITIES, FACTS AND RELATIONSHIPS



Поиск в графах

1. Поисковый запрос по ключевым словам Q
2. Найти список описаний именованных сущностей $E_1 \dots E_k$ из графовой базы знаний G , релевантных запросу
предположение – наличие релевантных фактов в базе знаний



Модель ранжирования – модель, оценивающая $P(E|Q)$

Рис. из презентации Н.Жильцова

Извлечение и анализ мнений

- Извлечение мнений
 - Выделение цитат
 - Классификация по источнику
 - Классификация по теме
 - анализ тональности
- Анализ тональности (sentiment analysis)
 - Анализ мнений о политиках, партиях
 - Извлечение и представление отзывов о товарах и услугах
 - Имидж компании

Извлечение мнений



Это какой-то ужас. В рецензии все описано так, что от фильма ждешь минимум продолжения фильма "Адреналина" с Стэтхэмом, но нет. 80 минут бесмысленных бегов, туда-сюда. Единтсвенный вопрос - Зачем?!?!



Хороший+, трешовый+ фильм, с отличным+ чувством юмора. Для любителей гая ритчи самое то, вот только картинка нищенская-, но ничего страшного+. Это даже колорит+ какой-то придает.



Я в дилом восторге+!!!!

Эта парочка неподражаема+

Так всё красиво+, аккуратно+,
технично+, с долей юмора, и ТАК
ЗАХВАТЫВАЮЩЩЩЩЩЩЩЕ
динамично+!!!!

В экстазе+)

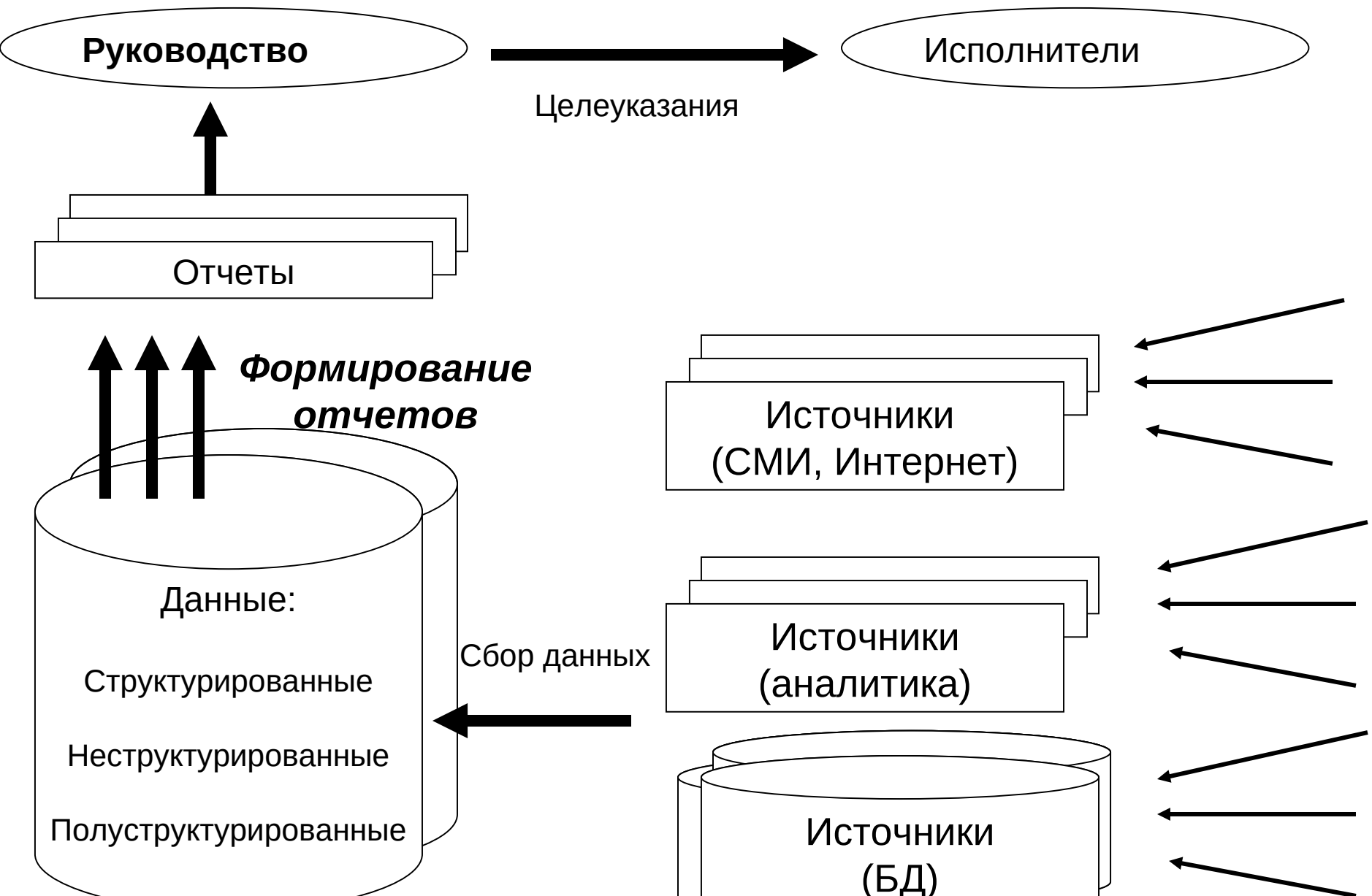
*Научно-исследовательский
вычислительный центр
МГУ имени М.В. Ломоносова*

*ООО «Лаборатория
информационных исследований»*

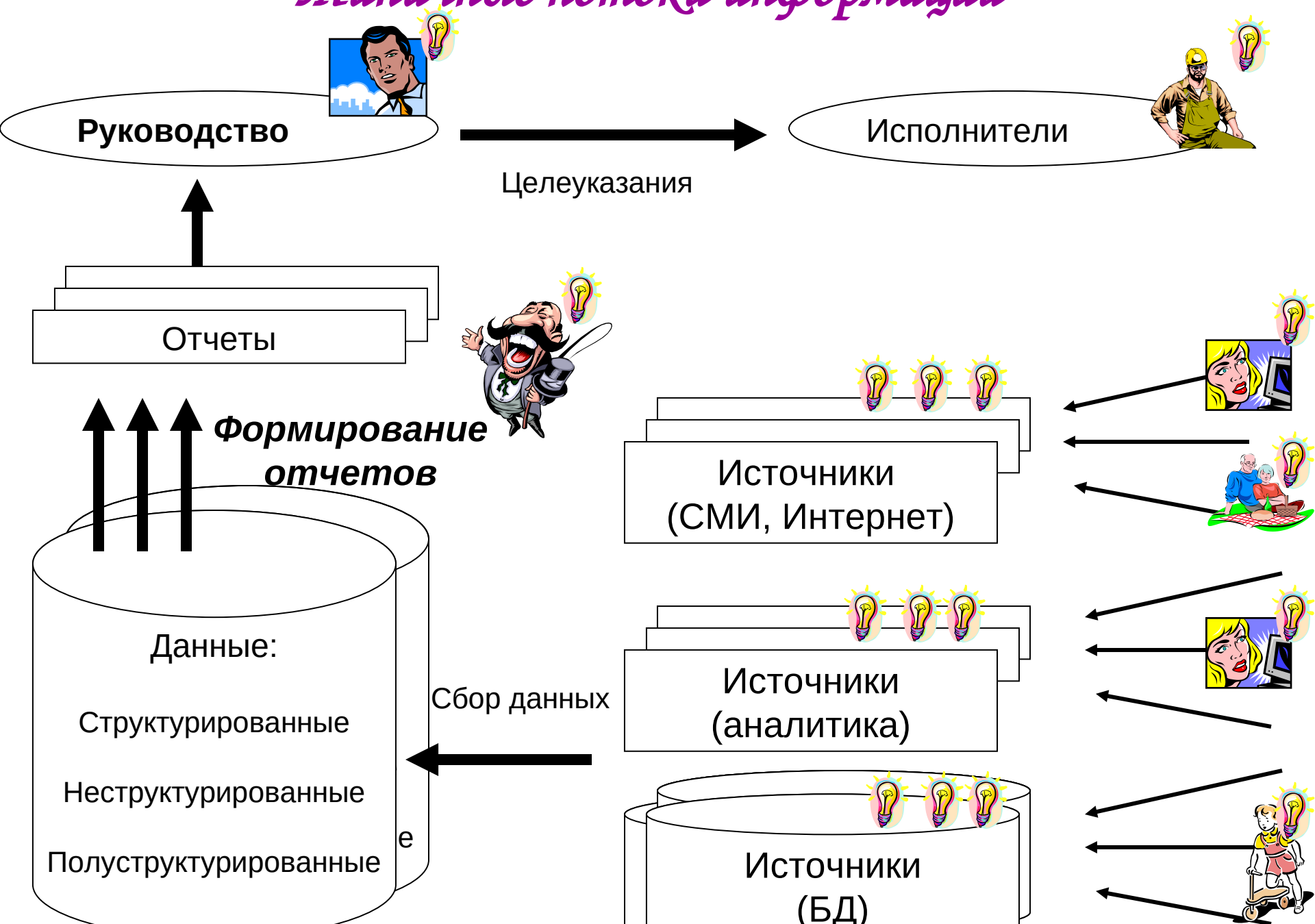
*Информационно-аналитические системы и
системы принятия решений*

о себе и нашей группе в МГУ

Типичные потоки информации в СТПР



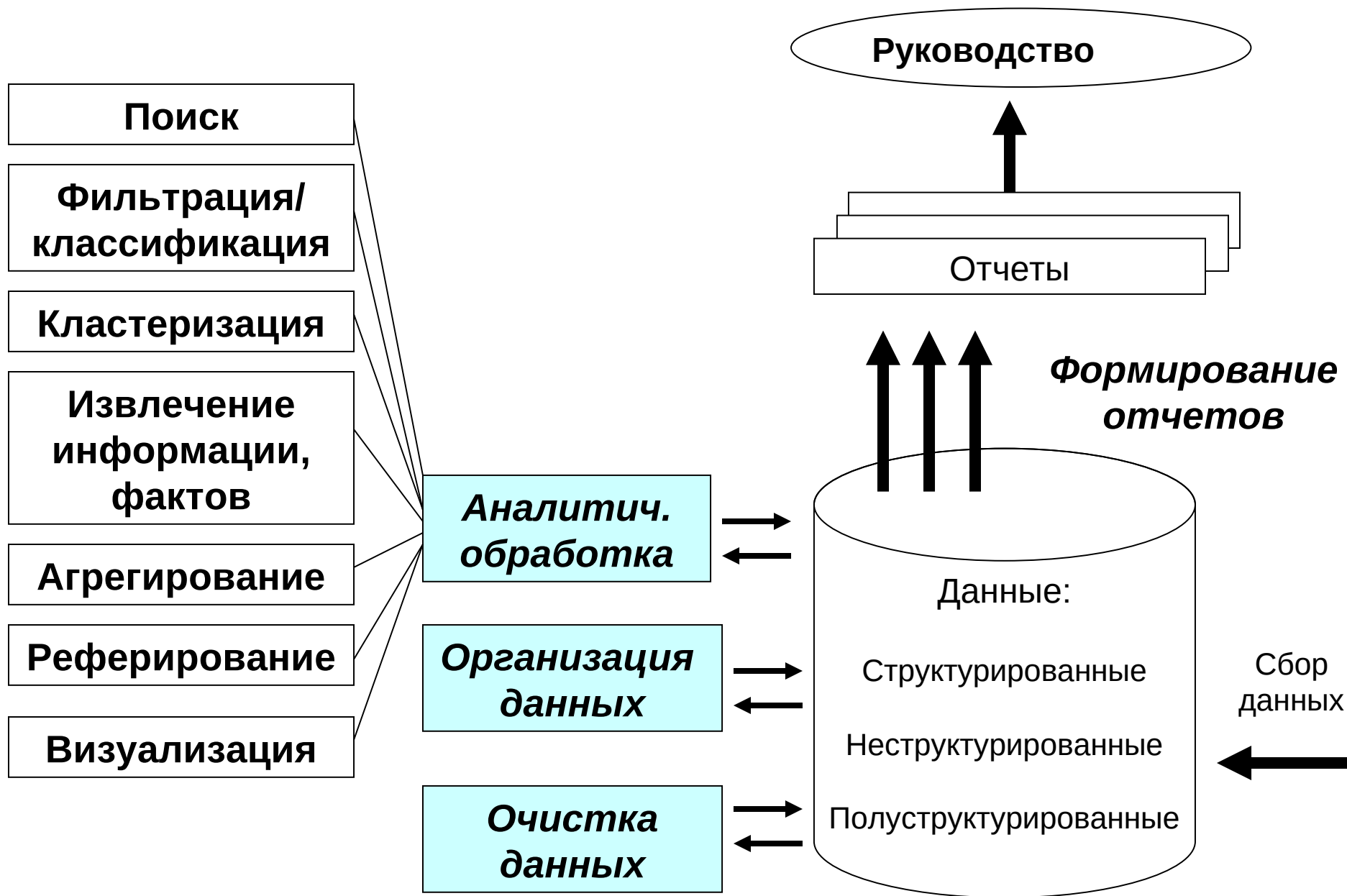
Типичные потоки информации



Задачи информационно аналитических систем

- **Situation awareness** (владение ситуацией, оценка обстановки)
 - напомнить, что происходило ранее
 - объяснить, что происходило, почему
 - мониторинг
 - объяснить, что сейчас происходит
- **Predictive analytics** (прогнозная аналитика)
 - проанализировать тенденции, тренды
 - экстраполировать ряды
 - ситуационное моделирование
 - мнения экспертов (форсайт)
- Представить результаты
 - отчет
 - визуализация

Внутренние потоки информации



Структура информационно-аналитической системы

**СИСТЕМА
СБОРА**
данных,
очистка и
конвер-
тация

**Лингви-
стико-
онтологи-
ческие
ресурсы**

словари,
словники,
тезаурусы,
таксо-
номии,
онтологии,
шаблоны

АЛОТ

фрагментация

морфология

терминология

тематический
анализ

рубрикация

аннотирование

сентимент

календарь

именованные
объекты

выделение
фактов

выделение
событий



БД

доку-
менты

мета-
данные

ПОды

словар
и
ЛО

сюжеты

мнения

клаузы

имена

факты

собы-
тия

класте-
ризация
доку-
ментов

группи-
рование
мнений

группи-
рование
клауз

группи-
рование
имен

группи-
рование
фактов

группи-
рование
событий

ИПС

поиск по
доку-
ментам

поиск по
кластерам
(сюжетам)

поиск по
мнениям

поиск по
клаузам

поиск по
именам

поиск по
фактам

поиск по
событиям

ИАС

ГИС

фасетный
анализ

времен-
ные ряды

OLAP

спектра-
льно-
фасетный
анализ

когни-
тивные
схемы

иссле-
дование
аналитики

интел-
лектуаль-
ные папки

ИАС+

анали-
тические
отчеты

корпора-
тивная
Вики-
педия

сценар-
ный
анализ и
прогно-
зирование

имитаци-
онное
модели-
рование

Знания о предметной области: Банковский тезаурус

Отношения на концептах

БАНКОВСК

Название концепта
БАНКОВСКАЯ КАРТА VISA ELECTRON
БАНКОВСКАЯ КАРТА VISA GOLD
БАНКОВСКАЯ КОМИССИЯ
БАНКОВСКАЯ ЛИЦЕНЗИЯ
БАНКОВСКАЯ МАРЖА
▶ БАНКОВСКАЯ ОПЕРАЦИЯ
БАНКОВСКАЯ ОПЕРАЦИЯ СО СЧЕТОМ

BANK TRANSACTION

Фильтр

Текстовый вход
▶ БАНКИНГ
БАНКОВСКАЯ ОПЕРАЦИЯ
БАНКОВСКАЯ ТРАНЗАКЦИЯ
КРЕДИТНО-РАСЧЕТНОЕ ОБСЛУЖИВАНИЕ
ОПЕРАЦИЯ БАНКА

Перейти к синонимам

Фрагменты текстов

Добавить

Изменить

Удалить

1 + -

Отношение	Аспект	Название концепта
ВЫШЕ		ФИНАНСОВАЯ ОПЕРАЦИЯ
НИЖЕ		АКТИВНАЯ БАНКОВСКАЯ ОПЕРАЦИЯ
НИЖЕ		БАНКОВСКАЯ ОПЕРАЦИЯ СО СЧЕТОМ
▶ НИЖЕ		БЕЗНАЛИЧНЫЙ РАСЧЕТ
НИЖЕ		ДЕПОЗИТНАЯ ОПЕРАЦИЯ
НИЖЕ		ДОВЕРИТЕЛЬНЫЕ ОПЕРАЦИИ
НИЖЕ		ДОКУМЕНТАРНАЯ ОПЕРАЦИЯ

CASHLESS SETTLEMENT

Текстовый вход
▶ БЕЗНАЛ
БЕЗНАЛИЧКА
БЕЗНАЛИЧНАЯ ОПЕРАЦИЯ
БЕЗНАЛИЧНАЯ ОПЛАТА
БЕЗНАЛИЧНАЯ СИСТЕМА
БЕЗНАЛИЧНАЯ ФОРМА
БЕЗНАЛИЧНАЯ ФОРМА РАСЧЕТОВ

Добавить

Изменить

Перейти

Удалить

1 + -

Добавить

Изменить

Удалить

1 + -

Закреть

Распознавание терминов в тексте

<input checked="" type="checkbox"/> ОПОЗНАННЫЕ ТЕКСТОВЫЕ ВХОДЫ И ИМЕННЫЕ СУЩЕ					Справка	Новая обработка
<input checked="" type="checkbox"/> Опознанные	<input type="checkbox"/> Поставить все	<input checked="" type="checkbox"/> Т	<input checked="" type="checkbox"/> Т_А	<input checked="" type="checkbox"/> Т_М		
	<input type="checkbox"/> ФИО	<input type="checkbox"/> Организации				
<input type="checkbox"/> Сантимент	<input type="checkbox"/> NECRF	<input checked="" type="checkbox"/> Подсвечивать фон				
<input type="checkbox"/> СПИСОК НАЙДЕННЫХ РУБРИК						
<input type="checkbox"/> СПИСОК НАЙДЕННЫХ РУБРИК САНТИМЕНТА						
<input type="checkbox"/> ТЕМАТИЧЕСКАЯ АННОТАЦИЯ						
<input type="checkbox"/> АННОТАЦИЯ						
<input checked="" type="checkbox"/> ОБРАБОТАННЫЙ ТЕКСТ						

Путин объяснил свои слова о чрезмерном укреплении рубля.

Заявление о рисках чрезмерного укрепления рубля не было попыткой повлиять на Центробанк, заявил президент Владимир Путин в интервью Bloomberg. По его словам, плавающий курс в долгосрочной перспективе сохранится

Сделанное в июле заявление о рисках чрезмерного укрепления рубля, которое участники рынка восприняли как сигнал о намерении властей сдержать усиление национальной валюты, не было попыткой повлиять на Центробанк, заявил президент Владимир Путин в интервью Bloomberg.

По его словам, Центробанк должен соответствовать уровню развития экономики. «Я поддерживаю контакт с властями, но никогда не даю им директивных указаний. И если я говорю, что рубль слишком сильно укрепился, я не говорю, что позиция Центробанка неправильная», — сказал Путин агентству.

В среду премьер-министром Дмитрием Медведевым обращил внимание на риски, связанные с ростом рубля, и предложил главе правительства подумать о дальнейших мерах в связи с этим. Незадолго до этого, в начале июля, эксперты Bloomberg назвали укрепление рубля помехой росту российской экономики: по их мнению, сильный рубль негативно сказывается на финансовых показателях российских компаний.

ПРЕЗИДЕНТ ГОСУДАРСТВА
 ПРЕЗИДЕНТ ОРГАНИЗАЦИИ
 Сантимент с аспектом - аспектный термин (O)
 Государственная власть
 Термин:
 T_M155308 ПРЕЗИДЕНТ ГОСУДАРСТВА
 M162349 ПРЕЗИДЕНТ ОРГАНИЗАЦИИ

Классификация текстов

ОПОЗНАННЫЕ ТЕКСТОВЫЕ ВХОДЫ И ИМЕННЫЕ СУЩЕ					Справка	Новая обработка
<input checked="" type="checkbox"/> Опознанные	<input type="checkbox"/> Поставить все	<input checked="" type="checkbox"/> Т	<input type="checkbox"/> T_A	<input type="checkbox"/> T_M		
	<input checked="" type="checkbox"/> ФИО	<input type="checkbox"/> Организации				
<input type="checkbox"/> Сантимент	<input type="checkbox"/> NECRF	<input checked="" type="checkbox"/> Подсвечивать фон				
<input checked="" type="checkbox"/> СПИСОК НАЙДЕННЫХ РУБРИК						
PRG Общий						
B010000000 Политика					93	
B030000000 Общество					90	
B020020000 Финансы					84	
B020000000 Экономика					84	
PRG Это профессиональный						
<input type="checkbox"/> СПИСОК НАЙДЕННЫХ РУБРИК САНТИМЕНТА						
<input type="checkbox"/> ТЕМАТИЧЕСКАЯ АННОТАЦИЯ						
<input type="checkbox"/> АННОТАЦИЯ						
▼ <input checked="" type="checkbox"/> ОБРАБОТАННЫЙ ТЕКСТ						
<p>Путин объяснил свои слова о чрезмерном укреплении рубля.</p> <p>Заявление о рисках чрезмерного укрепления рубля не было попыткой повлиять на Центробанк, заявил президент Владимир Путин в интервью Bloomberg. По его словам, плавающий курс в долгосрочной перспективе сохранится</p> <p>Сделанное в июле заявление о рисках чрезмерного укрепления рубля, которое участники рынка восприняли как сигнал о намерении властей сдержать усиление национальной валюты, не было попыткой повлиять на Центробанк, заявил президент Владимир Путин в интервью Bloomberg.</p> <p>По его словам, курс валюты должен соответствовать уровню развития экономики. «Я поддерживаю контакт с правлением и председателем правления ЦБ, но я никогда не даю им директивных указаний. И если я говорю, что рубль слишком окреп, я не говорю, что позиция Центробанка неправильная», — сказал Путин агентству.</p> <p>В середине июля президент на встрече с премьер-министром Сантимент с аспектом - оценочное слово (-1) мание на риски, подумать о дальнейших мерах в связи с этим. Незадолго до этого, в начале июля, эксперты Bloomberg назвали укрепление рубля помехой росту российской экономики: по их мнению, сильный рубль негативно сказывается на финансовых показателях российских компаний.</p>						

Извлечение имен

<input checked="" type="checkbox"/> ОПОЗНАННЫЕ ТЕКСТОВЫЕ ВХОДЫ И ИМЕННЫЕ СУЩЕ					Справка	Новая обработка
<input checked="" type="checkbox"/> Опознанные	<input type="checkbox"/> Поставить все	<input checked="" type="checkbox"/> Т	<input type="checkbox"/> Т_А	<input type="checkbox"/> Т_М		
	<input checked="" type="checkbox"/> ФИО	<input type="checkbox"/> Организации				
<input type="checkbox"/> Сантимент	<input type="checkbox"/> NECRF	<input checked="" type="checkbox"/> Подсвечивать фон				
<input type="checkbox"/> СПИСОК НАЙДЕННЫХ РУБРИК						
<input type="checkbox"/> СПИСОК НАЙДЕННЫХ РУБРИК САНТИМЕНТА						
<input type="checkbox"/> ТЕМАТИЧЕСКАЯ АННОТАЦИЯ						
<input type="checkbox"/> АННОТАЦИЯ						
▼ <input checked="" type="checkbox"/> ОБРАБОТАННЫЙ ТЕКСТ						

Путин объяснил свои слова о чрезмерном укреплении рубля.

Заявление о рисках чрезмерного укрепления рубля не было попыткой повлиять на Центробанк, заявил президент **Владимир Путин** в интервью Bloomberg. По его словам, плавающий курс в долгосрочной перспективе сохранится

Сделанное в июле заявление о рисках чрезмерного укрепления рубля, которое участники рынка восприняли как сигнал о намерении властей сдержать усиление национальной валюты, не было попыткой повлиять на Центробанк, заявил президент **Владимир Путин** в интервью Bloomberg.

По его словам, курс валюты должен соответствовать уровню развития экономики. «Я поддерживаю контакт с правлением и председателем правления ЦБ, но я никогда не даю им директивных указаний. И если я говорю, что рубль слишком окреп, я не говорю, что позиция Центробанка неправильная», — сказал **Путин** агентству.

В середине июля президент на встрече с премьер-министром **Дмитрием Медведевым** обратил внимание на риски, связанные с ростом рубля, и предложил главе правительства подумать о дальнейших мерах в связи с этим. Незадолго до этого, в начале июля, эксперты Bloomberg назвали укрепление рубля помехой росту российской экономики: по их мнению, сильный рубль негативно сказывается на финансовых показателях российских компаний.



/Дата_док="18.11.2013-21.11.2013"
/Термин_расш="ТРАНСПОРТНАЯ АВАРИЯ"
/Термин_расш="САМОЛЕТ"
/Термин_расш="КАЗАНЬ"

Поиск по персонам

Искать: ☐ Сообщения ☐ Сюжеты ☒ Имена ☐ Организации ☐ Результаты по датам ☐ Результаты по регионам ☐ Карта

[Расширенный поиск](#) --> |

Найдены 419 имен, 1007 документов. Показано, начиная с 1. (10 имен/стр.) ☐ убрать подсветку

1. [Ирек Минниханов](#) (235 /263 документов)

[Среди погибших в казанской авиакатастрофе оказался новосибирец](#) gorod54.ru 18.11.2013 21:37

[Самолеты "Аэрофлота" перевезут родственников погибших в авиакатастрофе в Казани](#) rus.ruvr.ru 19.11.2013 09:51

[Контейнер самописца разбившегося Boeing сильно поврежден](#) vz.ru 18.11.2013 12:56

2. [Рустам Минниханов](#) (111 /4246 документов)

[Губернатор Мурманской области выразила соболезнования президенту Татарстана](#) regions.ru 18.11.2013 14:45

[Среди погибших в казанской авиакатастрофе оказался новосибирец](#) gorod54.ru 18.11.2013 21:37

[Судмедэксперт: «Опознания погибших в авиакатастрофе в Казани не было»](#) izvestia.ru 19.11.2013 10:28

3. [Аксан Гиниятуллин](#) (75 /96 документов)

[Экипаж разбившегося Boeing впервые выполнял заход на второй круг](#) svpressa.ru 19.11.2013 11:49

[Мы приостановили эксплуатацию Boeing – глава авиакомпании "Татарстан"](#) ria.ru 19.11.2013 12:40

[Разбившийся в Казани Boeing должен был лететь в другой город](#) newizv.ru 19.11.2013 17:02

4. [Владимир Маркин](#) (69 /11078 документов)

[Речевой самописец с разбившегося в Казани "Боинга" найден поврежденным](#) rus.ruvr.ru 20.11.2013 18:34

[СК: Следствие располагает аудиозаписью переговоров диспетчера с экипажем Boeing](#) rbc.ru 19.11.2013 16:15

[Разбившийся в Казани oen-737 упал практически вертикально](#) nn.ru 18.11.2013 17:00

5. [Максим Соколов](#) (49 /1742 документов)

[Рассматривается пять версий крушения Boeing в Казани, теракт исключен](#) aif.ru 18.11.2013 14:37

[Компания «Татарстан» приостановила эксплуатацию Boeing 737](#) lenta.ru 18.11.2013 18:21

[Найдены бортовые самописцы Boeing 737, разбившегося в Казани](#) news.rufox.ru 18.11.2013 17:01

6. [Рустем Салихов](#) (46 /51 документов)

[Командир разбившегося Boeing ни разу не уходил на второй круг](#) vz.ru 19.11.2013 10:52

[Гендиректор "Татарстана": Погибшие пилоты впервые шли на второй круг при реальной посадке](#) vedomosti.ru 19.11.2013

[Командир разбившегося в Казани «Боинга 737-500» налетал на нем 510 часов](#) aif.ru 18.11.2013 14:03

7. [Александр Антонов](#) (36 /97 документов)

["Аэрофлот" бесплатно доставит родственников погибших в авиакатастрофе в Казань](#) amic.ru 19.11.2013 10:47

[Контейнер самописца разбившегося Boeing сильно поврежден](#) vz.ru 18.11.2013 12:56

Фасетный анализ

/Дата_док="01.08.2013-31.09.2013"
/Регион="АМУРСКАЯ ОБЛАСТЬ"
помощь

Искать

[Справка](#)

Искать: ☒ Сообщения ☐ Сюжеты ☐ Имена ☐ Организации ☐ Результаты по датам
☐ Результаты по регионам ☐ Карта | <- [Расширенный поиск](#) -> |

Найдено 2076 документов. Показано, начиная с 1. (10 док./стр.)

сортировать по релевантности

☐ убрать подсветку

Анализ:

--- Анализ по... ---

Кабмин выделил на [помощь](#) пострадавшим регионам Дальнего Востока 40 млрд руб. (86%)

2013-09-30 17:20:00.0000000 - [rbc.ru](#)
1744029

По Республике Саха пока нет точных данных, добавил он.

«Единая Россия» собрала почти 17 млн. для пострадавших от паводка (86%)

2013-08-31 09:03:00.0000000 - [er.ru](#)
1487256

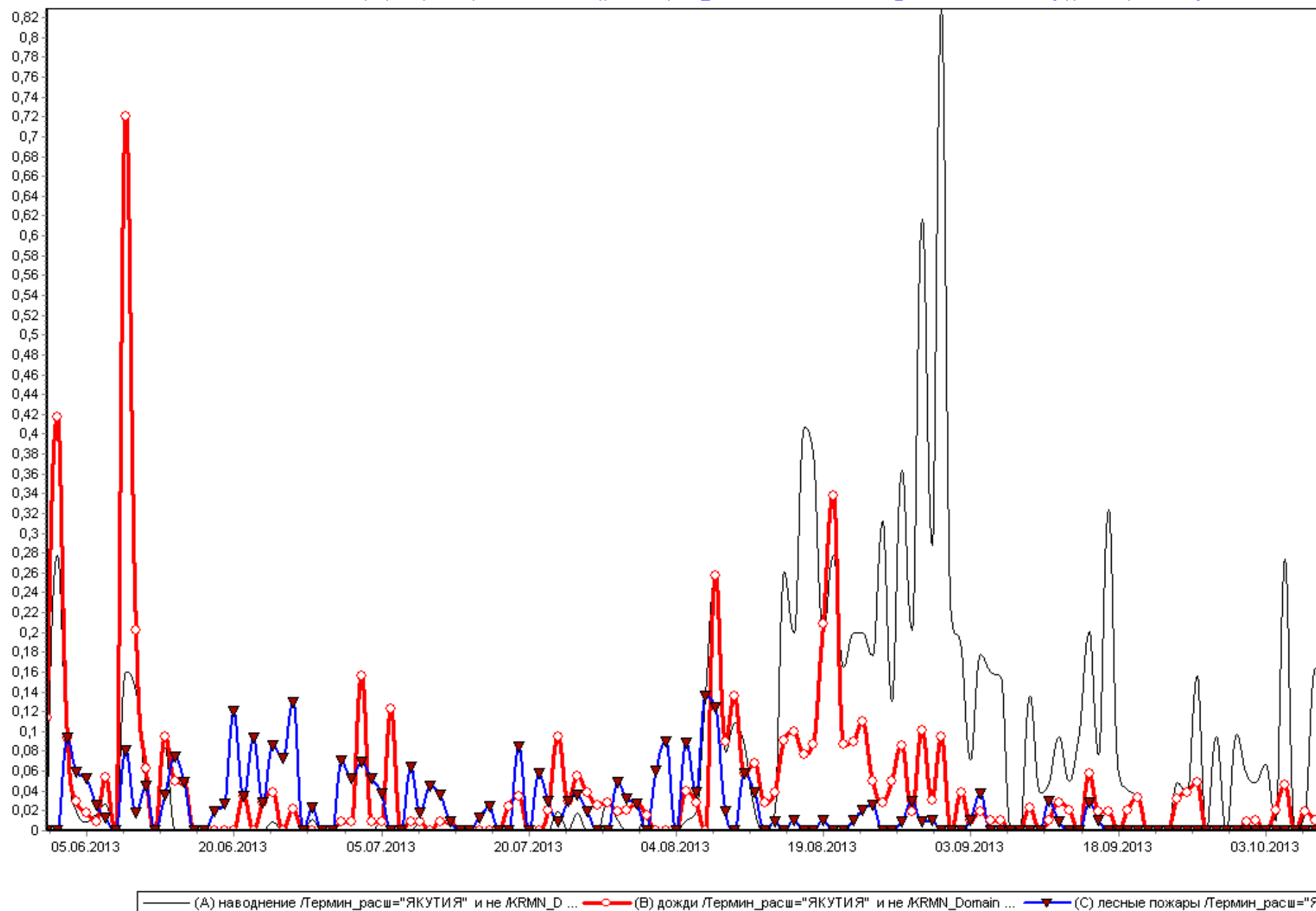
На повестке дня вопрос о [помощи](#) пострадавшим в результате паводка на Дальнем Востоке. Внимание высшего руководства страны к людям, оказавшимся в беде из-за стихии, вполне объяснимо: власть должна выполнить свою работу и защитить граждан.

Как сообщил руководитель проекта «Знак качества» Алексей Корягин, организации – партнеры партпроекта проявят солидарность с согражданами и окажут им возможную

+/-		Термин
+	+t	ПРИАМУРЬЕ
-	-t	
+	+t	БЛАГОВЕЩЕНСК
-	-t	
+	+t	АМУРСКАЯ ОБЛАСТЬ
-	-t	
+	+t	СОУЧАСТИЕ
-	-t	
+	+t	РАСЧЕТНЫЙ СЧЕТ
-	-t	
+	+t	ЕВРЕЙСКАЯ АВТОНОМНАЯ ОБЛАСТЬ
-	-t	
+	+t	АМУР
-	-t	

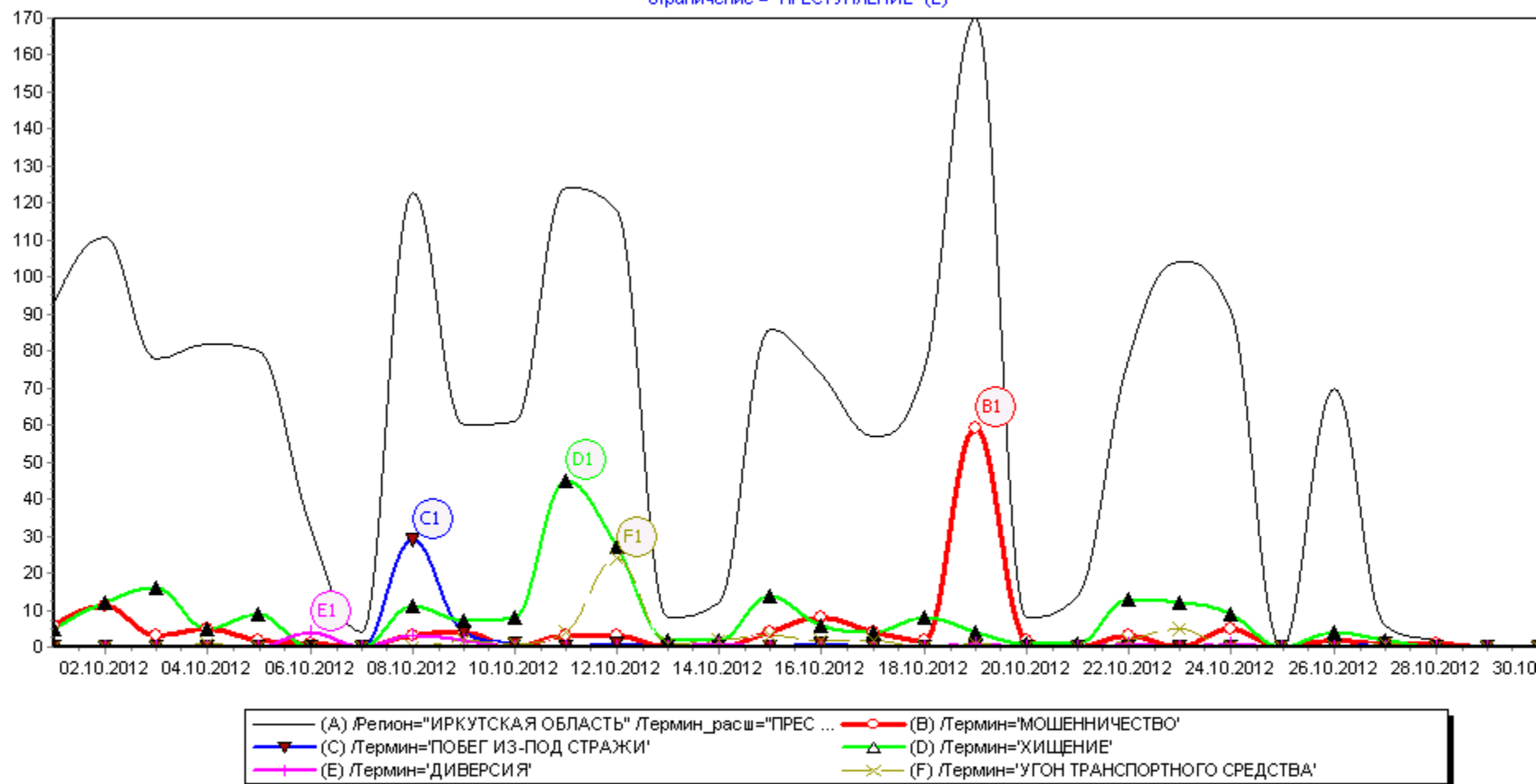
Якутия: лесные пожары vs. дожди vs. наводнения

Процент публикаций по теме == наводнение /Термин_расш="ЯКУТИЯ" и не /KRMN_Domain="rus.ruvr.ru" == [БД=Default, rank>-100]



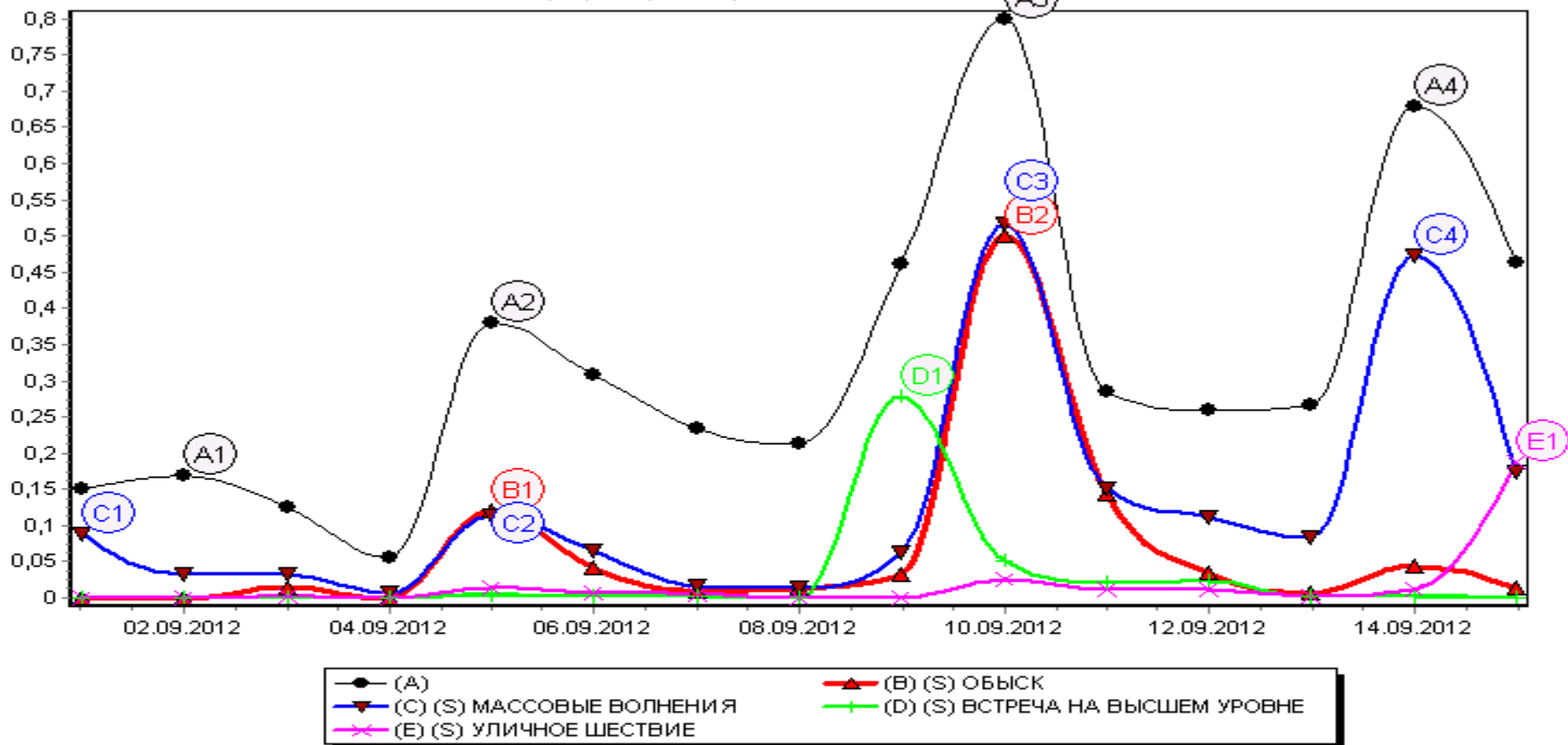
Спектрально-фасетный анализ преступлений в Иркутской области по видам

Ограничение = ПРЕСТУПЛЕНИЕ (L)



B1 19.10.2012 В Москве по подозрению в покушении на мошенничество задержан экс-чиновник из Иркутской области //quebmvd.ru
 B1 19.10.2012 11:49:00 Главу департамента ТНК-BP подозревают в продаже госпостов //ОРБК
 B1 19.10.2012 11:56:00 Задержан глава департамента по взаимодействию с госорганами "ТНК-BP" //ОГолос России - новости
 C1 08.10.2012 10:21:00 Трое заключенных совершили побег из колонии №96 в Иркутске //1Per_Газета Иркутск
 C1 08.10.2012 14:29:00 Сбежавшие заключенные пойманы в пригороде Иркутска //1Per_interfax-russia.ru Сибирь
 C1 08.10.2012 14:33:00 Пойманы заключенные, сбежавшие из колонии в Иркутске //Вечерняя Москва
 D1 11.10.2012 4:41:00 Трое подростков в Приангарье подозреваются в нападении на почтальона //1Per_АиФ Иркутск новости

Процент публикаций по теме == Собчак ==



C4 14.09.2012 11:18:00 Ксения Собчак раскритиковала награждение Pussy Riot премией лучший арт-проект года //ИД "Собеседник"

C4 14.09.2012 14:47:00 Ксении Собчак не хватает скандальной популярности Pussy Riot //Allnews4.me

C4 14.09.2012 15:18:00 Ксения Собчак оценила награду Pussy Riot //Firstnews

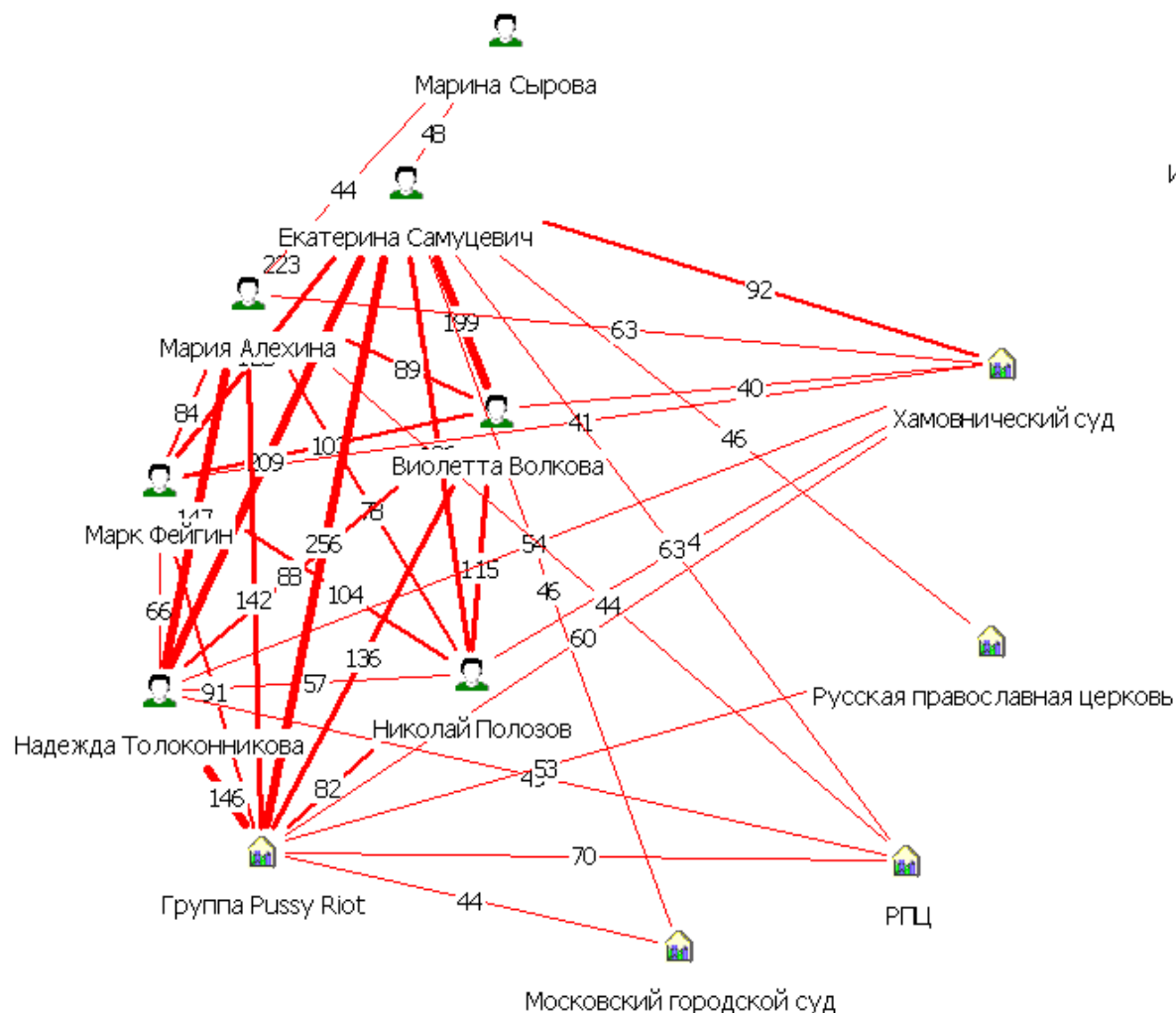
D1 09.09.2012 16:05:00 09.09.2012 16 05 Ксения Собчак была немало удивлена тем, что запись в ее твиттере обсуждал президент Владим

D1 09.09.2012 16:41:00 Собчак удивилась, когда Путину на саммите АТЭС рассказали о ее Twitter //ONEWSru.com

D1 09.09.2012 18:11:00 "Мне кажется, Владимир Владимирович ответил как настоящий альфа-журавль, поэтому мне его ответ понрав

E1 15.09.2012 9:34:00 На «Марш миллионов» идут тысячи //BREM.RU - лента новостей бизнеса

Стандартная когнитивная схема: именованные объекты (люди- организации)



Интерфакс

МВД

ПАСЕ

Совет ЕЕ

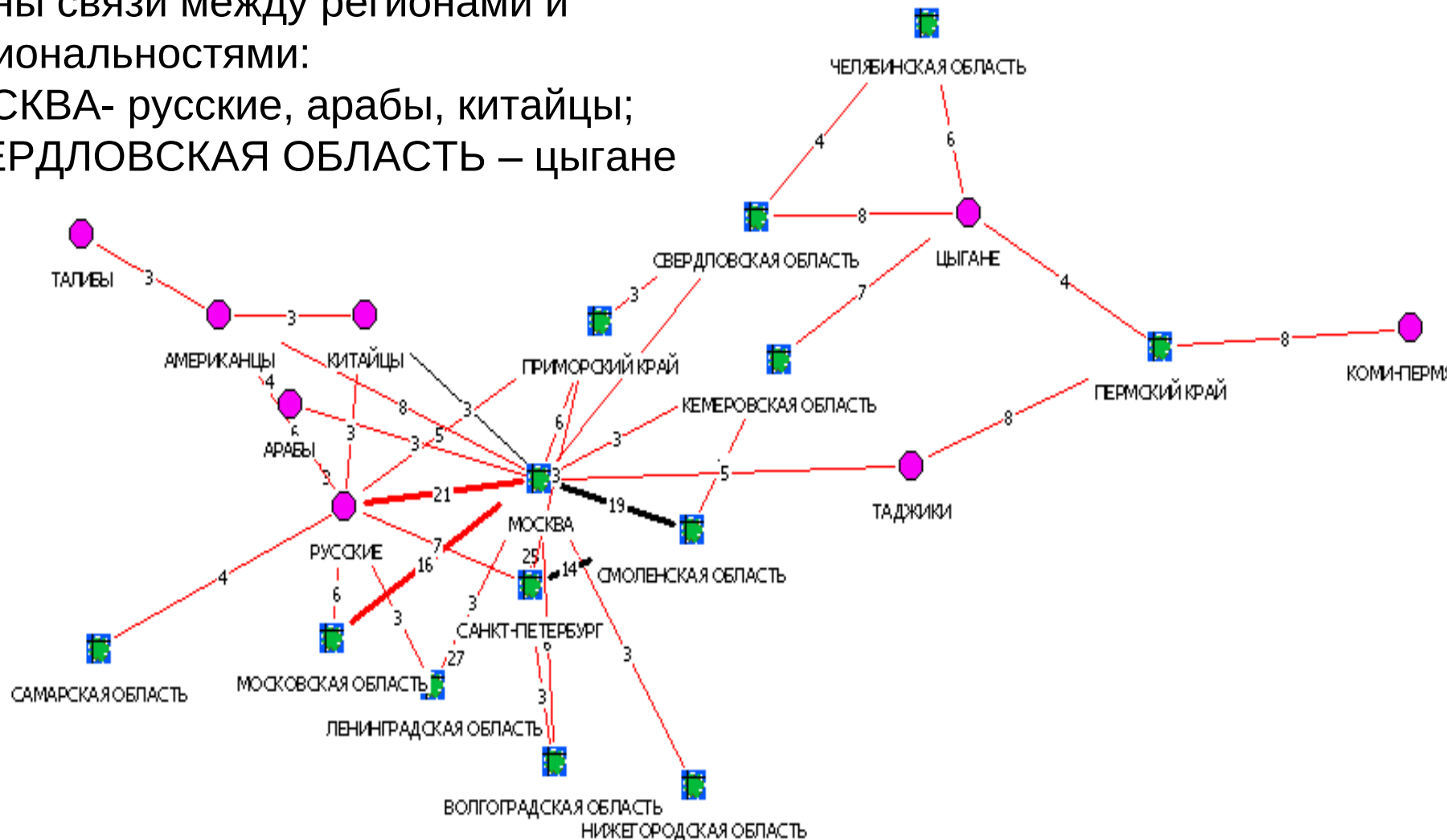
Добавление в когнитивную схему иерархий понятий

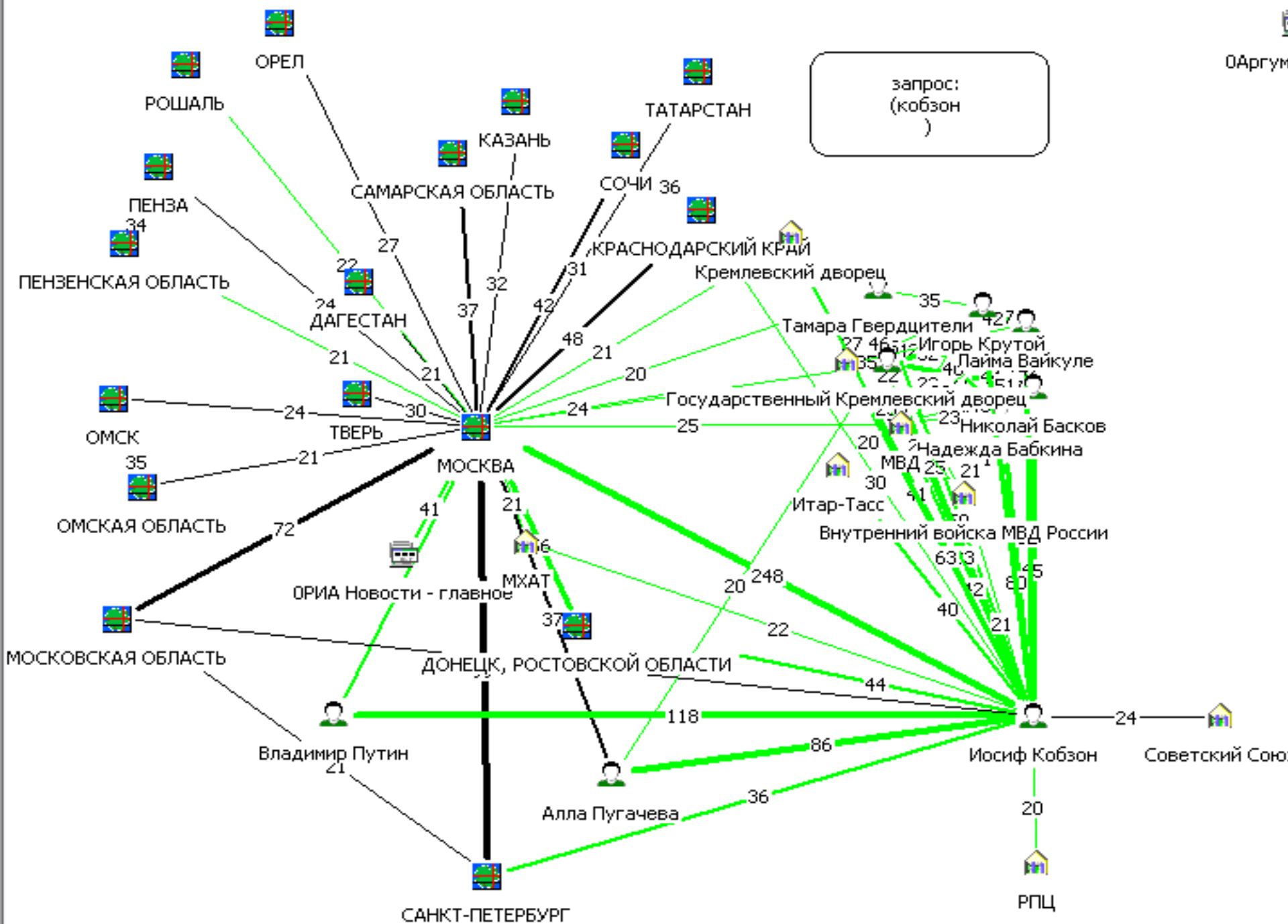
Запрос: «героин /Дата_док="09.2012"»

видны связи между регионами и национальностями:

МОСКВА- русские, арабы, китайцы;

СВЕРДЛОВСКАЯ ОБЛАСТЬ – цыгане





Наши проекты

Проекты в области обработки нормативно-правовых документов

Нормативно-правовые акты

```
graph TD; A[Нормативно-правовые акты] --> B[УИС РОССИЯ, 1995-2002-...]; A --> C[«Юпитер» ИПИ РАН, 1996-1997]; A --> D[Минюст РФ Армада [Эр-Си-О], 2004]; A --> E[Система терминологии Счетная палата РФ, 2003]; A --> F[ФКЗ «Право» ГАС «Выборы» [НИИ «Восход»] 1997-2012]; A --> G[«Гарант», 2000-2012]; A --> H[ИПС «Стенограммы» ГД РФ ФС 1997-н/в]; G --> I[Классификация]; G --> J[Вопросы-ответы]; G --> K[Аннотирование, сравнение судебных решений]
```

УИС РОССИЯ,
1995-2002-...

«Юпитер»
ИПИ РАН,
1996-1997

Минюст РФ
Армада [Эр-Си-О],
2004

Система терминологии
Счетная палата РФ,
2003

ФКЗ «Право»
ГАС «Выборы»
[НИИ «Восход»]
1997-2012

«Гарант»,
2000-2012

Классификация

Вопросы-ответы

Аннотирование, сравнение
судебных решений

ИПС
«Стенограммы»
ГД РФ ФС
1997-н/в

«Изумруд»
УОПИ ФСО РФ,
[«Кодекс»]
2007-2008

Проекты в области обработки новостных документов и СМИ

Новости и СМИ

```
graph TD; A[Новости и СМИ] --> B[УИС РОССИЯ, 1995-2002-...]; A --> C[«Юпитер» ИПИ РАН, 1996-1997]; A --> D[Сайты по инфобезопасности НИЦ «Квант», 2004]; A --> E[Рамблер.Новости 2008-2013]; A --> F[ГосЗаказчик]; F --> G[Классификация]; F --> H[Аннотирование]; F --> I[Аналитические отчеты]; F --> J[Календарь событий];
```

УИС РОССИЯ,
1995-2002-...

«Юпитер»
ИПИ РАН,
1996-1997

Сайты по
инфобезопасности
НИЦ «Квант»,
2004

Рамблер.Новости
2008-2013

ГосЗаказчик

КП «Новости»
ЦБ РФ [Эр-Си-О]
2006-н/в

Классификация

Аннотирование

Аналитические отчеты

Календарь
событий

Проекты в области оценки тональности

Тональность (сентимент)

```
graph TD; Root[Тональность (сентимент)] --> DE[Dialogue.Evaluation 2013-2016]; Root --> GZ[ГосЗаказчик]; Root --> KH[«Киноход» 2011]; Root --> Y[Яндекс 2014]; Root --> RN[Рамблер.Новости 2008-2013]; Root --> CK[Чистка комментариев]; DE --> B[Блоги]; DE --> TW[Твиттер]; DE --> T[Товары]; DE --> TE[Телеком]; DE --> BK[Банки]; Y --> R[Рестораны]; Y --> KF[Кинофильмы]; Y --> MP[Мобильные приложения]; Y --> CK;
```

Dialogue.Evaluation
2013-2016

Блоги

Твиттер

Товары

Телеком

Банки

Рамблер.Новости
2008-2013

Чистка
комментариев

ГосЗаказчик

«Киноход»
2011

Яндекс
2014

Рестораны

Кинофильмы

Мобильные приложения

Проекты в области анализа и визуализации данных

Анализ и визуализация данных

```
graph TD; A[Анализ и визуализация данных] --- B[Фасетный анализ поисковой выдачи УИС РОССИЯ, 1995-2002-...]; A --- C[Мониторинг градостроительства «Гранит-Центр», 2005-2006]; A --- D[Поисковая выдача на картограммах]; A --- E[Когнитивные схемы]; A --- F[Анализ временных рядов]; A --- G[Концептуальные схемы в образовании [Фонд Бортника] 2017]; A --- H[ГосЗаказчик];
```

Фасетный анализ
поисковой выдачи
УИС РОССИЯ,
1995-2002-...

Мониторинг
градостроительства
«Гранит-Центр»,
2005-2006

Поисковая
выдача на
картограммах

Когнитивные
схемы

Анализ
временных рядов

Концептуальные схемы
в образовании
[Фонд Бортника]
2017

ГосЗаказчик

Литература

- Manning, Shutze. Introduction to Information Retrieval
– <http://nlp.stanford.edu/IR-book/>
- К. Маннинг, П. Рахгаван, Х. Шютце Введение в информационный поиск. Изд-во Вильямс, 2011
- Дополнительно:
- Большакова Е. И., Воронцов К. В., Ефремова Н. Э., Клышинский Э. С., Лукашевич Н. В., Сапин А. С. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие. М. : НИУ ВШЭ, 2017.
– https://miem.hse.ru/clschool/the_book
- Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011
– <https://publications.hse.ru/mirror/pubs/share/folder/gcd5r3mn96/direct/50492942>
- Лукашевич Н.В. Тезаурусы в задачах информационного поиска. Изд-во Моск. ун-та, 2011.
– http://www.labinform.ru/pub/ruthes/book/louk_book.pdf

Домашнее задание

- Заглавная страница Википедии
- Раздел: Знаете ли Вы?
- (Из новых статей Википедии)
- **Итальянский спортсмен** в один год стал чемпионом страны по футболу и олимпийским призёром в фехтовании.
- Есть архив

Домашнее задание (продолжение)

- В архиве найти неделю, когда у вас день рождения (2020-2021 год)
 - Взять три факта
 - И найти соответствующие предложения, содержащие этот факт в статьях Википедии
 - Записать в файл
 - Сделать выводы, легко ли их бы было найти автоматически.
- Легко – это когда
 - Все слова запроса в одном предложении
 - Большое количество совпадающих слов

Отчет присылаем на почту

- vmk_ir@mail.ru