

01-02- Приложения автоматической обработки текстов

Прикладные задачи: машинный перевод

Начало исследований - 50ые годы 20го века

- Джоржтаунский эксперимент, 54 г.: автоматический перевод с русского на английский, словарь – 250 слов
- Первые работы в России: 55 г. – перевод с английского на русский текстов по ПМ, словарь – 2300 слов; далее – работы в ИПМ имени Келдыша (О.С. Кулагина)
- Простейшая лингвист. модель: *пословный* перевод
- Неравномерность развития работ по МП (приостановка финансирования исследований в 60-е годы)
- Периодизация методов/систем: используемая лингвистическая стратегия и лингвист. ресурсы
- 50-60 гг. – двуязычные системы,
пословный и пословно-пооборотный перевод
(приемлемое качество только для родственных языков, например: испанский-португальский)

Машинный перевод: Поколения систем

- 60-70 гг. – *пофразный* перевод, синтаксический анализ, стратегия АНАЛИЗ \Rightarrow ТРАНСФЕР \Rightarrow СИНТЕЗ
 - модульность (грамматика и словарь)
 - пред- и пост-редактирование человеком
 - появление промышленных систем:
SYSTRAN – США, 70 г. , перевод научно-техн. текстов
- 70-80 гг., экстенсивное развитие – *многоязычные* системы, идея внутреннего универсального семантического языка-посредника (для европейских языков)
 - ИнформЭлектро / ИППИ – система ЭТАП, основана на модели ЕЯ «Смысл \Leftrightarrow Текст» , французско/английско-русский перевод научно-технических текстов
- 80-90 гг. – многоязычные системы,
 - опора на лексические и терминологические БД
 - использование *интерлингвы* – языка-посредника
 - система ЭТАП-3 – язык *UNL*
- 90-2000 гг. – использование статистики, корпусов текстов: *статистическая трансляция* (переводчик Google, позже Яндекс)

Извлечение информации (*Information extraction*)

Извлечение информации (знаний) из текстов:

- Специфика задачи – распознавание и выявление в текстовой коллекции релевантной информации:
 - конкретных объектов (имен лиц, названий фирм и т.п)
 - отношений (связей) выделенных объектов и понятий
 - связанных с ними событий и фактов
 - понятий (терминов), напр.: *технология двойной накачки*
- Методы извлечения, основанные на правилах:
 - *частичный синтаксический анализ* текстов - *shallow approach*
 - *лингвистические шаблоны*, содержащие лексическую, морфологическую и синтаксическую информацию
- Итеративная разработка правил и шаблонов

Пример извлечения информации

- «Краткосрочный государственный кредит в размере \$4 миллиарда получит компания Chrysler»
- Типы извлекаемой информации:
 - имена сущностей
 - отношения между сущностями
 - факты

Фрейм: Факт получения кредита

Слоты	Значения
Сумма	\$4 миллиарда
Получатель	Chrysler

Извлечение терминов

- Терминологические слова и словосочетания – называют понятия специальной области знаний:
общий регистр, число с плавающей точкой, пенсионное обеспечение
- Извлечение терминов и связей терминов (*род-вид, часть-целое*)
- Критерии выделения:
 - статистические (частотность)
 - лингвистические (шаблоны):
например: шаблоны определений терминов
- Приложения:
 - построение *гlossариев* и *предметных указателей*
 - создание *онтологий* и *тезаурусов* ПО (*моделей* ПО)
 - поддержка терминологического анализа текстов
 - навигация по терминам текста

Извлечение и анализ мнений

- Близко по целям и методам *к направлению Information Extraction*
- *Opinion Mining* – извлечение и анализ мнений, отзывов (о персоналиях, товарах, услугах, фильмах, книгах и проч.) - из сети Интернет (форумы, блоги и т.п.)
- *Sentiment Analysis* – анализ тональности текстов (положительная, отрицательная, нейтральная)

Извлечение мнений



Well as usual Keanu Reeves is nothing special–, but surprisingly, the very talented+ Laurence Fishbourne is not good– either, I hope they do not shoot a sequel.



Good+ film with an excellent+ sense of humor. For fans of Guy Ritchie. Only a picture is poor–.



The actors are first grade+ and it has a really well thought out story line. I've seen it 10 times and I'll watch it a few more. Enjoy!

Writing support

Автоматизация подготовки и редактирования текстов

- Первые программы:
 - автоматическая постановка переносов слов
 - проверка орфографии (спеллеры, автокорректоры)
- Коммерческие системы:
проверка орфографии ,
частично синтаксиса, а также – сложности стиля
- Исследовательские разработки:
 - выявление неправильного употребления предлогов (использование моделей управления)
 - обнаружение сложных лексических ошибок: описки, приводящие к другим словам: *овальный/оральный*; паронимические ошибки: *болотный/болотистый*

Информационный поиск

- Индексирование документа на ЕЯ – выделение ключевых слов и словосочетаний
 - критерии автоматического индексирования:
статистические и лингвистические
- Классификация текстов – отнесение к классам с заданными свойствами/параметрами
- Рубрицирование текстов – классификация, соотнесение с иерархической системой классов
- Кластеризация текстов – создание подмножеств тематически близких документов
- Реферирование текста – построение краткого реферата для одного или нескольких текстов
- Аннотирование текста - краткое описание содержания текста (простейший случай – список ключевых слов)

Информационный поиск

автоматическая обработка текстов — Яндекс: Нашлось 12 млн ответов - Mozilla Firefox

Файл Правка Вид Журнал Закладки Инструменты Справка

http://yandex.ru/yandsearch?text=автоматическая+обработка+текстов&lr=213

Самые популярные (Без имени) (Без имени) (Без имени)

louk_nat@mail.ru: Re: Your IEKA 201... автоматическая обработка те... YTP.O.ru, новости дня pogoda.rbc.ru - Погода Прогноз пог...

Поиск Почта Карты Маркет Новости Словари Блоги Видео Картинки ещё




Яндекс
Нашлось 12 млн ответов

автоматическая обработка текстов

☐ в найденном ☐ в Москве расширенный поиск

Найти

Мои находки
Настройка
Регион: Москва


-  [АОТ :: Главная](#)
Автоматическая Обработка Текста.
aot.ru
-  [Введение в лингвистику и автоматическая обработка текстов, часть 1](#)
1. Введение в теоретическую лингвистику. Свойства и функции естественного языка. Знаковая природа языка. Типы знаков. Язык, речь, речевая деятельность.
shad.yandex.ru
-  [Категория:Автоматическая обработка текстов — Википедия](#)
Страницы в категории «Автоматическая обработка текстов» Показано 7 страниц этой категории из 7.
ru.wikipedia.org
- [УИС РОССИЯ : Технология автоматической обработки текстов](#)

Яндекс Директ

[RCO: компьютерная лингвистика](#)
технологии интеллектуальной обработки данных на естественном языке
[Адрес и телефон](#) www.rco.ru

[Разместить объявление по запросу «автоматическая ...»](#)

«автоматическая ...» в картинках



Все картинки

Готово

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

Проблемы интернет-поиска

- На некоторые запросы много релевантных документов – как выбрать лучшие
- На некоторые запросы мало документов – как найти
- Спам
- Неоднозначные запросы: линкольн
- Диверсификация выдачи

Диверсификация выдачи поисковой системы

Яндекс
Нашлось
27 млн ответов

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

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

[Мои находки](#)
[Настройка](#)
Регион: Москва

Результаты **все** [в рунете](#) [в мировом интернете](#)

1

n

[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

n

[shop.nokia.ru/](#)
[shop.nokia.ru](#)

3

N

[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

W


[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»](#)

Nokia Lumia 920 против iPhone 5.
Сравнение AppleInsider.ru **Nokia** Lumia..
[Все видеоролики](#)

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

Автоматическая рубрикация (text categorization)

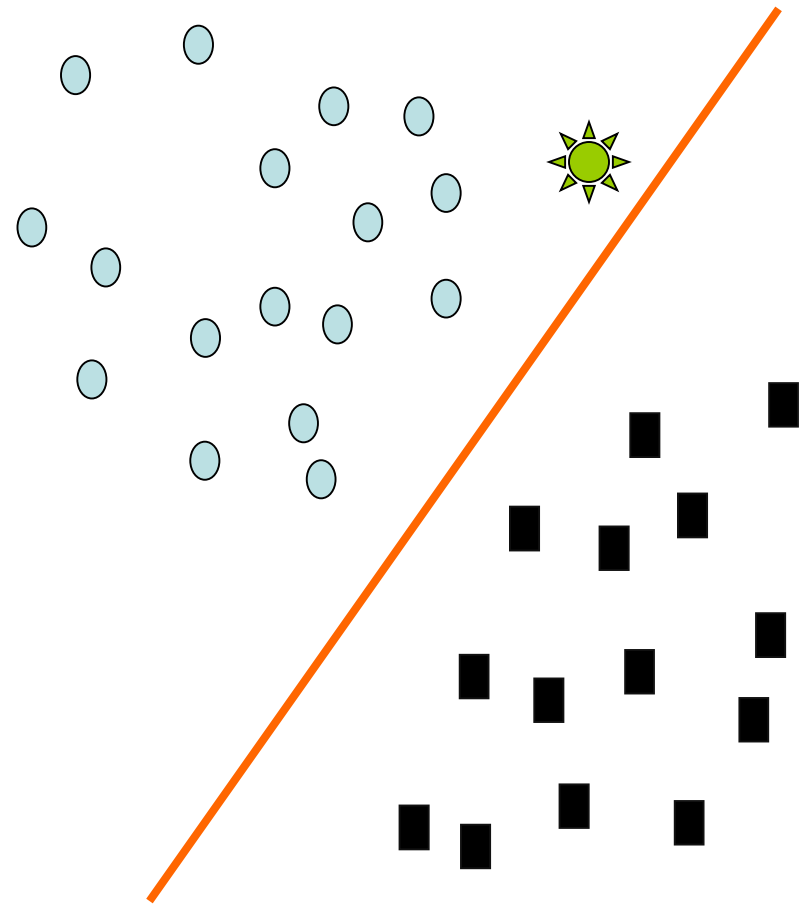
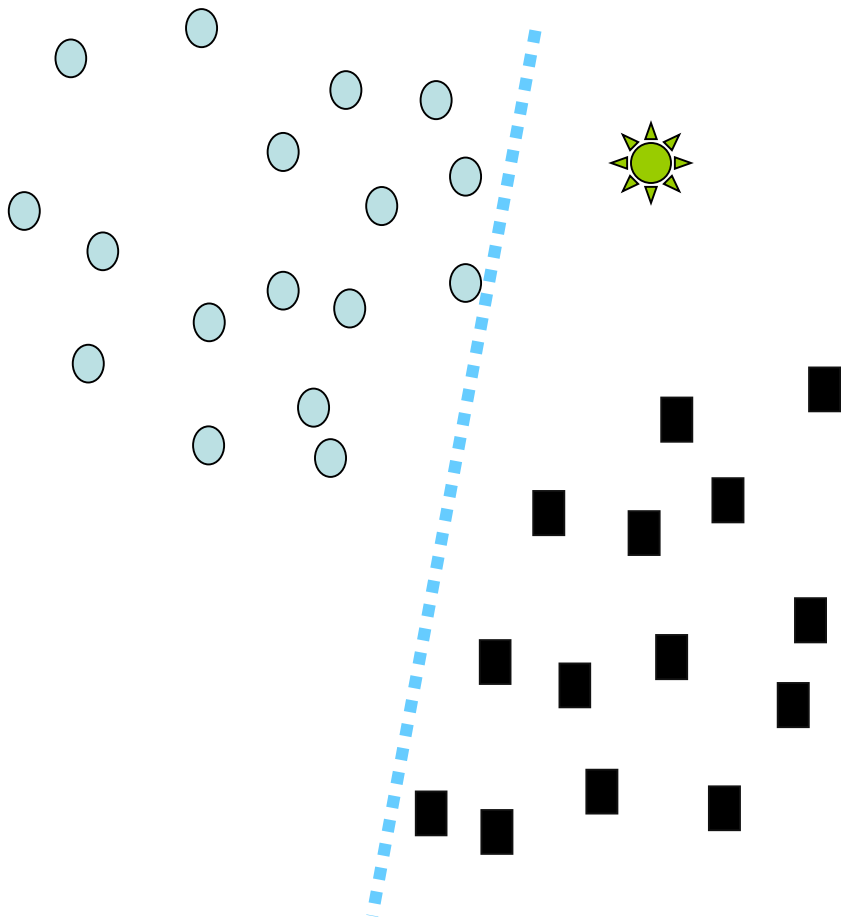
- Имеется заранее определенный набор рубрик (categories) – рубрикатор
- Нужно приписать документам текстового массива (потока) набор рубрик возможно с весами соответствия
- Рубрикаторы
 - количество рубрик от нескольких до тысяч
 - с иерархией или без нее

Примеры рубрикаторов

- Каталог Интернет-сайтов:
Open Directory Project – dmoz.org
 - 4,830,584 sites, 75,151 editors, over 590,000 categories

- C
 - [Arts](#)
[Movies](#), [Television](#), [Music](#)...
 - [Business](#)
[Jobs](#), [Real Estate](#), [Investing](#)...
 - [Computers](#)
[Internet](#), [Software](#), [Hardware](#)...
 - [Games](#)
[Video Games](#), [RPGs](#), [Gambling](#)...
 - [Health](#)
[Fitness](#), [Medicine](#), [Alternative](#)...
 - [Home](#)
[Family](#), [Consumers](#), [Cooking](#)...
 - [Kids and Teens](#)
[Arts](#), [School Time](#), [Teen Life](#)...
 - [News](#)
[Media](#), [Newspapers](#), [Weather](#)...
 - [Recreation](#)
[Travel](#), [Food](#), [Outdoors](#), [Humor](#)...
 - [Reference](#)
[Maps](#), [Education](#), [Libraries](#)...
 - [Regional](#)
[US](#), [Canada](#), [UK](#), [Europe](#)...
 - [Science](#)
[Biology](#), [Psychology](#), [Physics](#)...
 - [Shopping](#)
[Autos](#), [Clothing](#), [Gifts](#)...
 - [Society](#)
[People](#), [Religion](#), [Issues](#)...
 - [Sports](#)
[Baseball](#), [Soccer](#), [Basketball](#)...
 - [World](#)
[Deutsch](#), [Español](#), [Français](#), [Italiano](#), [Japanese](#), [Nederlands](#), [Polska](#), [Dansk](#), [Svenska](#)...

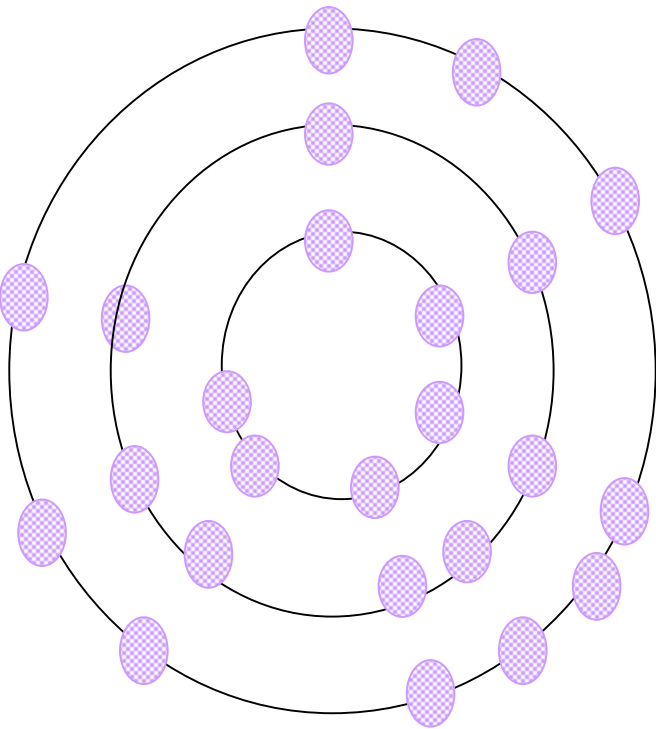
Положительные и отрицательные примеры: как лучше отделить



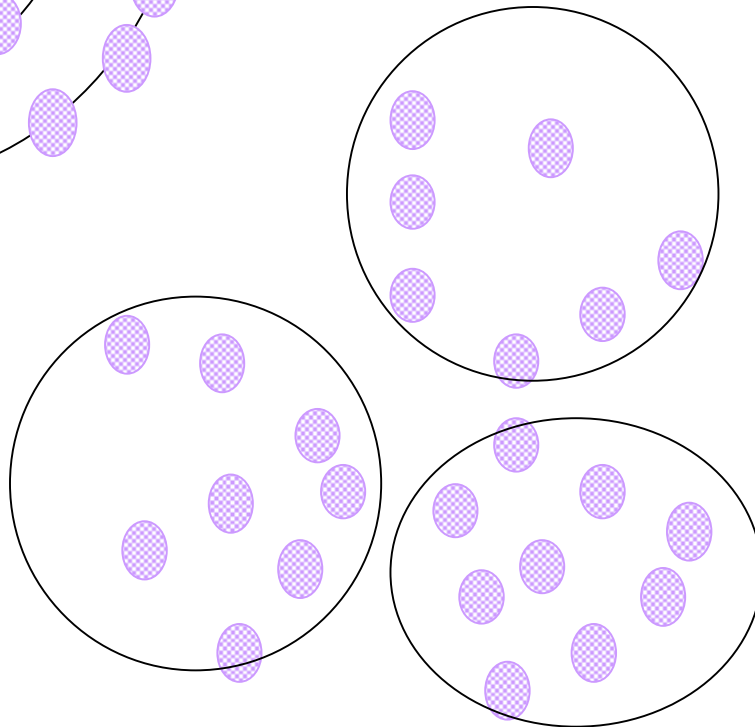
Автоматическая кластеризация текстов

- 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)

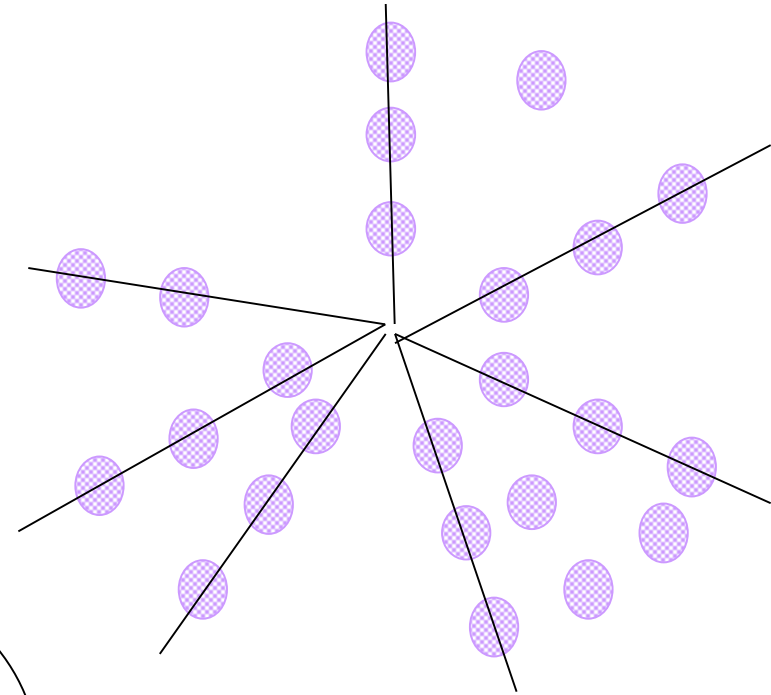
Варианты кластеризации



**кластеризация
по общим
вопросам**



**сюжетная
кластеризация**



**кластеризация
по объектам**

Новостные агрегаторы





Яндекс
Поиск Погода Карты Маршруты Новости Словари Видео Фото Курсы Книжки Ещё ▼
Войти

Найти

Пользователи ищут: африканская чума свиней ▼ расширенный поиск

Главные новости
Политика
В мире
Общество
Экономика
Спорт
Присшествия
Культура
Наука
Здоровье
Hi-Tech
Интернет
Авто
Туризм

Автоматически обработано 2641 источник,
обновлено в 14:16 мск

Выпуск: Россия | Украина


▲ новость часа

В жилищном департаменте Москвы опровергают проведение обысков (54 сообщения)


Как сообщили ИТАР-ТАСС в пресс-службе Главного управления МВД РФ по Центральному федеральному округу, обыски проводятся в рамках ранее возбужденного уголовного дела по статье УК РФ "Мошенничество".

Россия и страны Персидского залива решили отозвать нефть от доллара (81)

Страны Персидского залива вместе с Китаем, Россией, Японией и Францией разрабатывают план отказа от долларов в торговле нефтью.


Медведев назначил начальником своей референтуры Еву Василевскую (62)  13 мнений

Как говорится в сообщении кремлевской пресс-службы, согласно распределению полномочий в администрации главы государства, работу референтуры курирует помощник президента Джанхан Поллыева, которая работала спичрайтером у президентов Бориса Ельцина и Владимира Путина.

США будут вести диалог с КНДР при условии отказа от ядерной программы (92) 

Ким Чен Ун занял пост, равный по значимости должности заместителя руководителя партии, сообщает 6 октября агентство Associated Press со ссылкой на представителя Южной Кореи Сонг Сонг-Гван.

Новости Москвы



- Суд подтвердил отказ выплатить 5 млн рублей пострадавшему от Бескова
- Генерал МВД пострадал в ДТП в Москве
- Расследование уголовного дела в отношении майора Бескова завершено

Все новости Москвы >>

Другие регионы >>

Платите
за квартиру
из квартиры

Вход | Карта | Видео | Фото | Новости | Группы | Email | еще ▼

Google Россия

Новости

Персональная версия - **Главны́е новости** **Поиск**

- Главны́е новости
- В мире
- Россия
- Бизнес
- Наука и техника
- Экология Подмосковья
- Sci/Tech (U.S.)
- Спорт
- Культура
- Здоровье

> Любое содержание
Заголовки
Изображения


Поиск новостей Поиск в Интернете Расширенный поиск новостей Настройки

Обновление 6 мин. назад

В деле об убийстве Анны Политковской появились новые фигуранты

Радиостанция ЭХО МОСКВЫ • 18 мин. назад

Спустя три года после убийства журналистки Анны Политковской, дети считают, что следствию дан новый и последний шанс найти заказчиков и убийц. Время уходит и надежда на раскрытие дела все меньше, сказала Вера Политковская сегодня на пресс-конференции. Напомним, что ранее Верховный суд вернул...



Newsinfo

Жилдепартамент Москвы объясняет в связи с делом о мошенничестве Lenta.ru • 33 мин. назад - все статьи (67) »

Нобелевскую премию по физике получили ученые из США и Англии

Утро.Ru • 22 мин. назад - все статьи (47) »

Сообщения, которые планируют освещать РИА Новости 6 октября

РИА Новости • 05.10.2009 - все статьи (9) »

Communication pioneers win 2009 physics Nobel

Reuters • 10 мин. назад - все статьи (820) »

Кадыров против Орлова: заседание продолжается

Радио Свобода • 15 минута назад - все статьи (92) »

Монумент «Работорг» и колхозники: вернут на место 5 декабря

Полит.ру • 2 часа: назад - все статьи (20) »

В России началась промышленная набортка вакцин против свиного гриппа

Интерфакс • 48 Мин. назад - все статьи (260) »

Корейский депутат рассказал, когда Ким Чен Ир откажется от власти

Росбалт RU • 1 час назад - все статьи (339) »


Как студенты СФУ прошли практику на Саяно-Шушенской ГЭС

Комсомольская правда • 57 мин. назад

Саяно-Шушенская ГЭС стала не только местом, где произошла страшная авария, но и своеобразным учебным полигоном. Саяно-Шушенская ГЭС стала не только местом, где произошла страшная авария, но и своеобразным учебным полигоном. студенты Саяно-Шушенского филиала Сибирского Федерального Университета


Братская нагрузка Саяно-Шушенской ГЭС Газета Ру Комиссия проверит готовность СУГЭС к строительным работам имидж РИА Новости IA REGNUM - Утро Ру - Росбалт RU - Российская Газета Все похожие статьи: 1 292 »

Суд отклонил жалобу одного из



RIA Самара

Дмитрий Медведев Юрий Лужков Владимир Путин Кристина Орбакайте Гус Хиддинк Кирилл Лухов Петр Сумин



[Интернет](#)
[Новости](#)
[Картинки](#)
[Видео](#)
[Top100](#)
[Товары](#)
[Визитки](#)
[еще ▾](#)

[Выпуск России](#) | [Украина](#)

[ГЛАВНОЕ](#)
[КАРТИНА ДНЯ](#)
[КАРТИНА НЕДЕЛИ](#)

[РОССИЯ](#)
[МИР](#)
[БИЗНЕС](#)
[РЫНКИ](#)
[ТЕХНОЛОГИИ](#)
[НАУКА](#)
[СПОРТ](#)
[КИНО](#)
[КУЛЬТУРА](#)
[МУЗЫКА](#)
[ЖИЗНЬ](#)
[АВТО](#)
[ЖЕНСКИЙ КЛУБ](#)
[ЗДОРОВЬЕ](#)
[ЗВЕРИНА И СЕКС](#)

ГЛАВНОЕ

Инфляция в РФ остается нулевой второй месяц подряд

Инфляция в России в сентябре была нулевой. Об этом говорится в сообщении Росстата. За период с января по сентябрь 2009 года инфляция составила 0,1 процента.

Вы заглянули рост цен на продукты в последние месяцы? Проголосуйте

- Верховный суд сократил срок убийце Анны Бельской
- В деле Анны Политковской появились «новые лица»
- Дума поддержит выделение бюджетных средств АвтоВАЗу
- Японцы создали работавший на метаноле спиртелефон
- Хиндики раскрыл преималыне сборной
- Суд подтвердил отказ в компенсации пострадавшему от Енисоя
- Правительство сократит армию киноинко
- Старукини МВД прелели обыски в департаменте химииной политике Москвы
- Монголии отплате Канаде обещанное России нсторонение

РОССИЯ

Бензин подешевел

Суд принял законное отпаше каретировать брах москвичек

В Санкт-Петербурге прелев ночной иторм

Виктор Ерофеев отпегил на «форменный доносо» филологаш

«Рабочего и колонизинк» вернут на место 5 декабря

МИР

Власти Китая решили создать крупнейше мадиноконини

француские субмарини онасат протипотоплендыни системами

Сторонники Януковича разблорковали Веруюую Раду

Хункороеский депутат навалел срок уооада Ким Чен Ира

Индонезия поспитан на вооружение российских ракеты

БИЗНЕС

«Лукойл» и Казахстан поспали французам на Каспий

В США «обанкротились» сразу три банка

Societe Generale увеличинилат капитал на 4,8 миллиарда евро

ВБ притискут к созданию коодинной госиперини

РЫНКИ




9 октября закончинется прием заявок на XII Конкурс годовых отчетов

Российские фондовые торги отпрыхлись ростом

Доллар упал ниже 30 рублей

Торги в Азии прорудат разнонаправленно

ПОГОДА: МОСКВА

 +7°
  +2°
  +8°

[Дни](#)
[Ночи](#)
[Утра](#)

КУРСЫ ВАЛЮТ

ЦБ (66.0)	Почта	Продажа
\$ 30,8785	2970	30,85
€ 40,0259	43,70	44,10

Наши новостные каналы

[illegible]

Автоматическое аннотирование

- Индикативная аннотация — пересказ основного содержания текста
- Контекстно-зависимая аннотация
 - сниппеты в поисковых системах
- Аннотация многих документов
 - Аннотация новостных кластеров
- Основной метод:
 - выделение наиболее информативных предложений
- Проблемы?

Аннотирование новостного кластера в 2013 году

Это не
начало
текста

Яндекс **НОВОСТИ** Поиск Почта Карты Маркет Новости Словари Блоги Видео Картинки ещё Войти Помощь

Найти расширенный поиск

☐ только в этом сюжете

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

Подать заявку

История систем АОТ

- Конец 50-х годов – машинный перевод
- Составная часть исследований в область искусственного интеллекта
- 60-е – 80-е годы языковые модели, правила, грамматики
- 70-е нужны знания о предметной области, АОТ-системы в конкретных предметных областях
 - All grammars leak
- 80-е годы поиск адаптивных механизмов, позволяющих настраиваться системы АОТ на новые предметные области и тексты

История систем АОТ-2

- 90-е годы по н/в
- Стало много текстов
- Использование статистических методов
- Автоматическая обработка текстов (как и Искусственный интеллект в целом) стала экспериментальной наукой
 - Тщательную экспериментальную проверку (тестирование) предложенного подхода

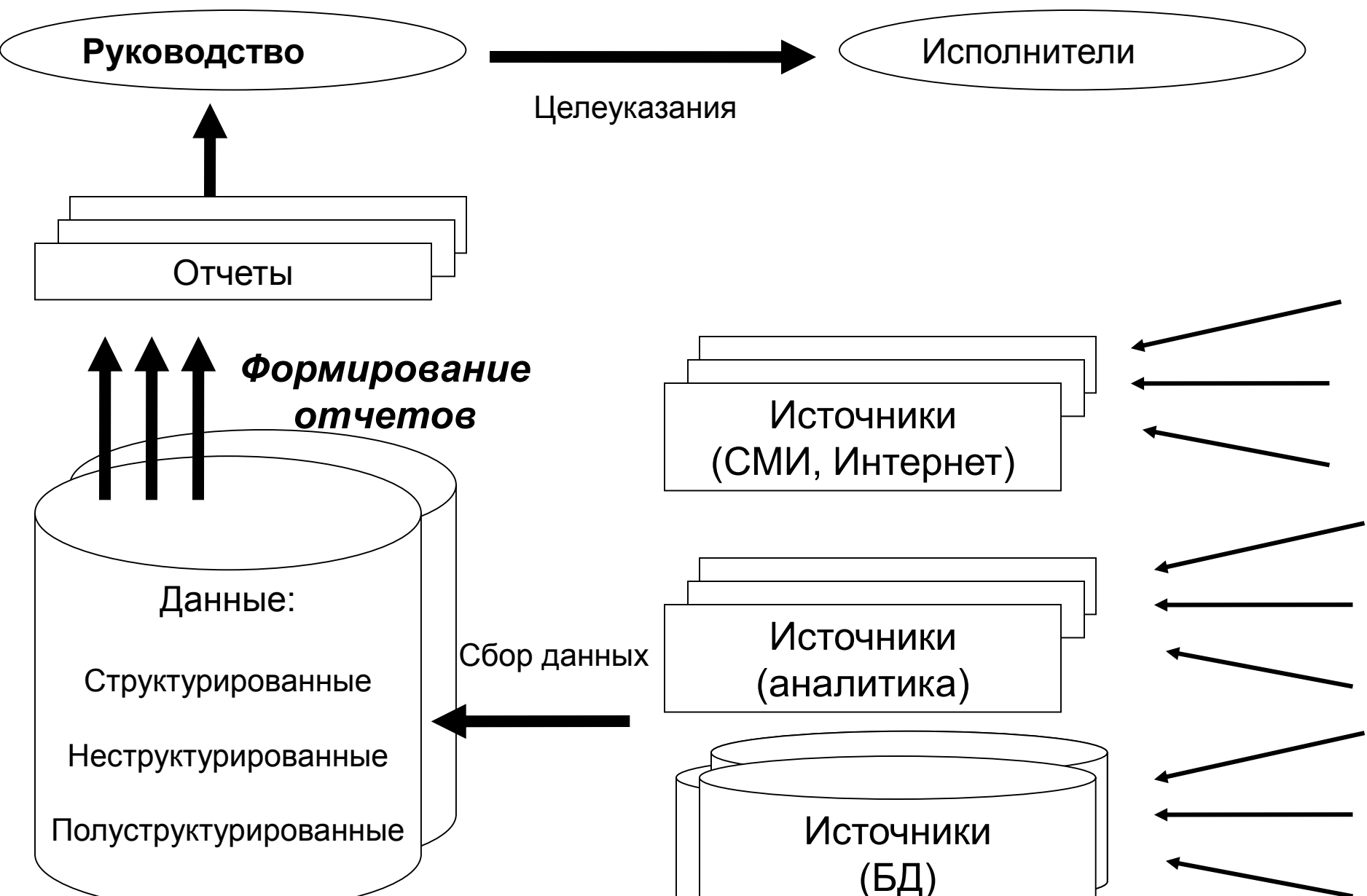
Программа курса

- Сентябрь:
 - введение, статистический анализ, морфологический анализ, словосочетания
- Октябрь
 - Информационный поиск и приложения
- Ноябрь
 - Синтаксис, семантика
- Декабрь
 - Различные приложения

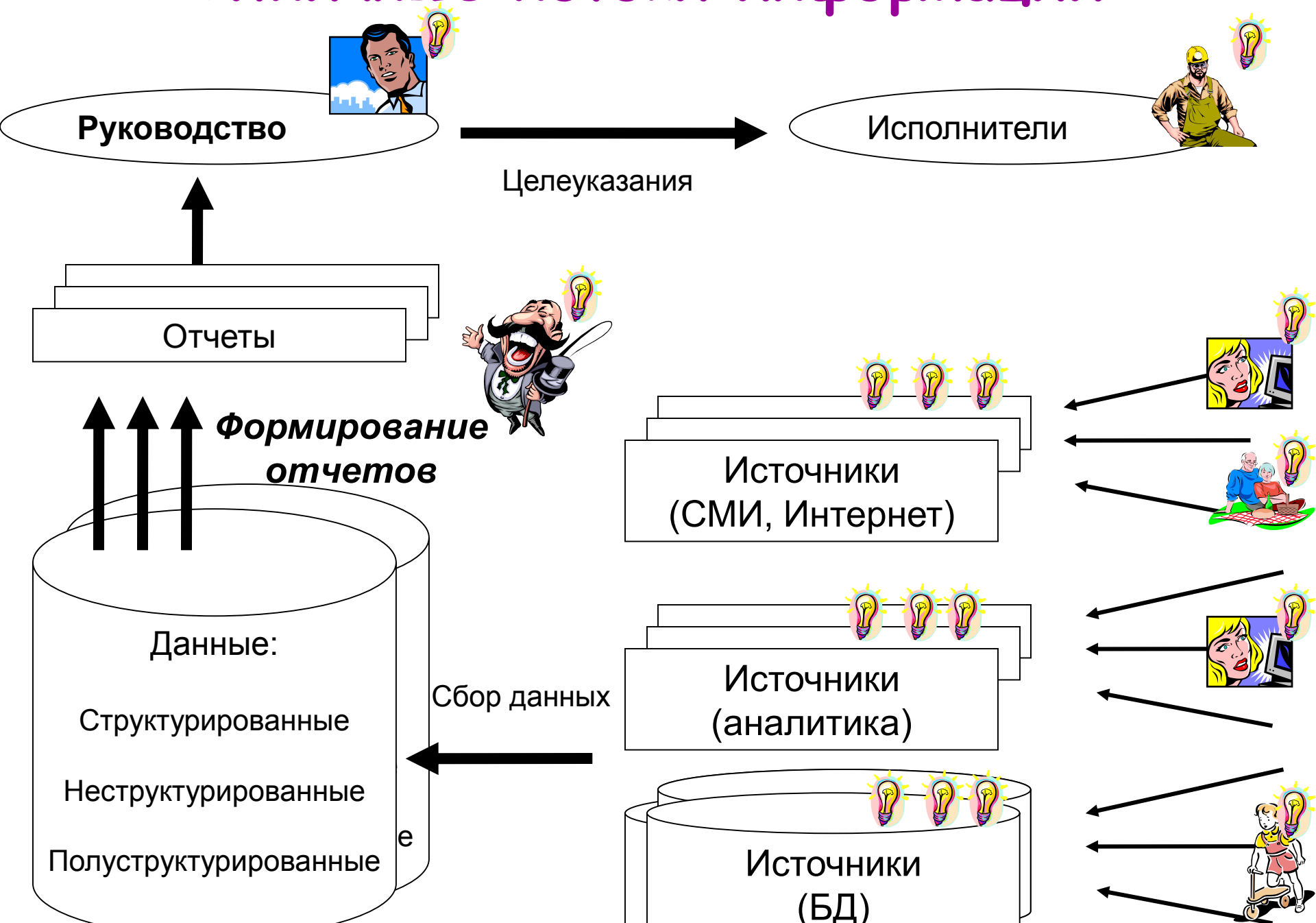
Информационно-аналитические системы и системы принятия решений

о себе и нашей группе в МГУ

Типичные потоки информации в СТТР



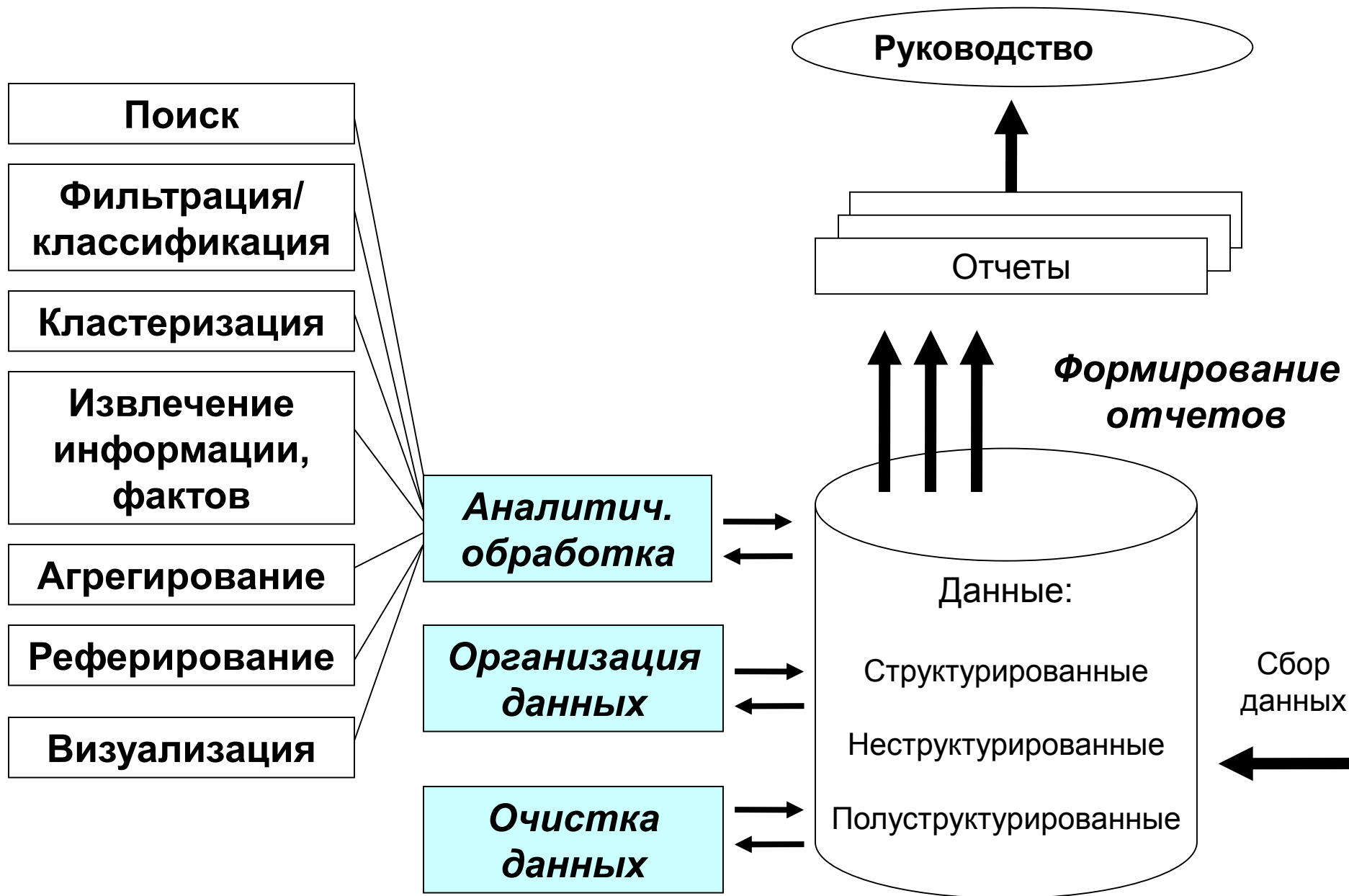
Типичные потоки информации



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+

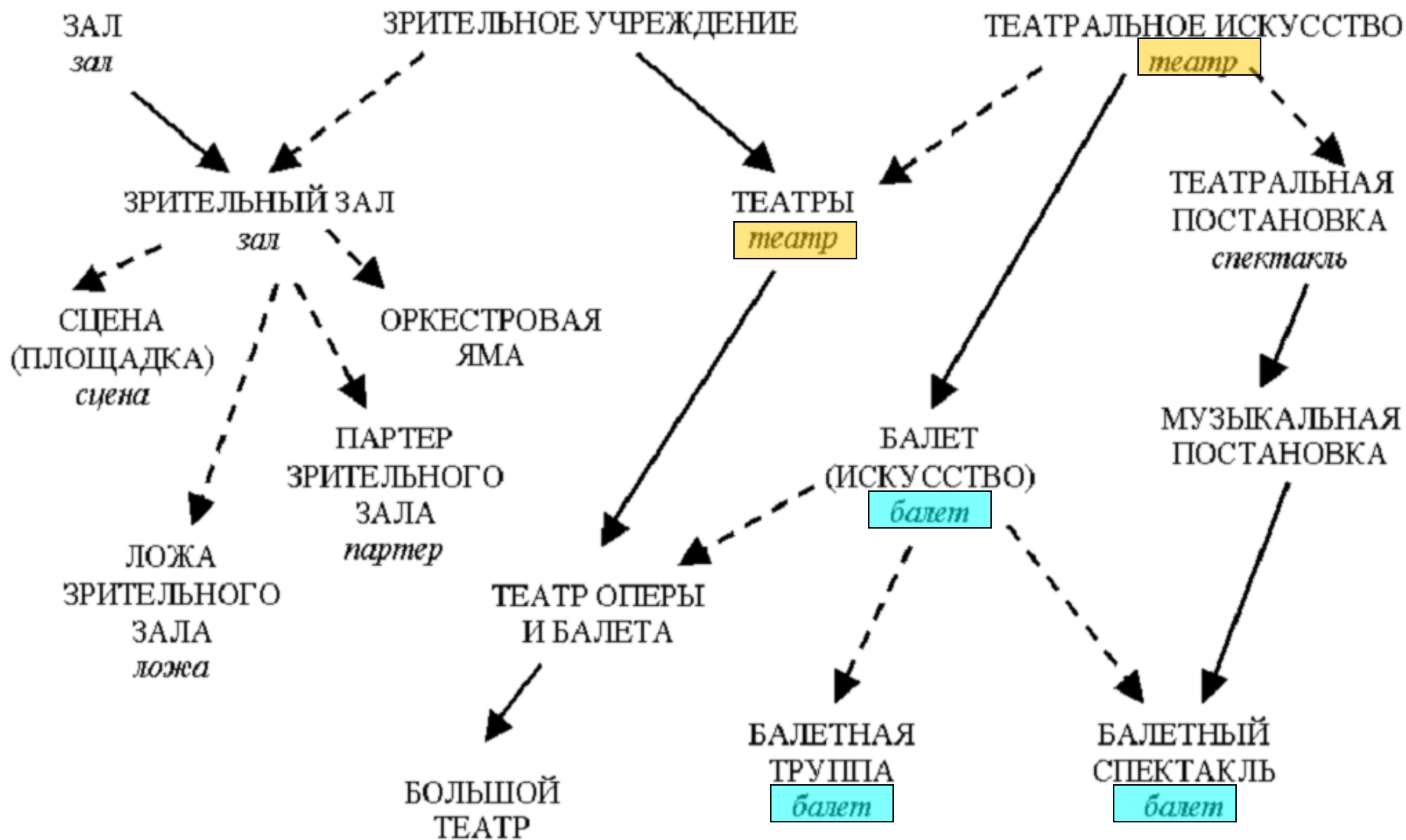
анали-
тические
отчеты

корпора-
тивная
Вики-
педия

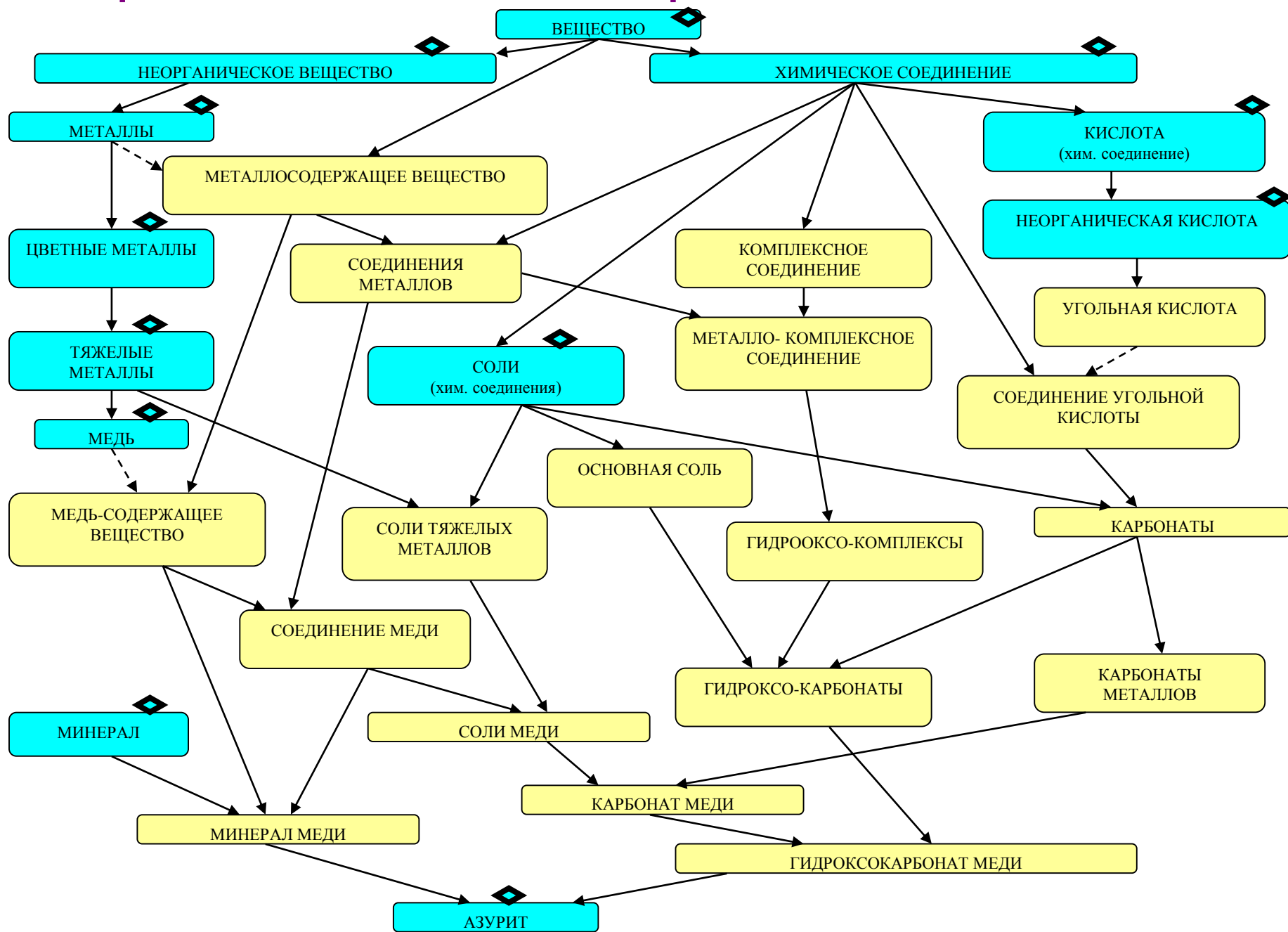
сценар-
ный
анализ и
прогно-
зирование

имитаци-
онное
модели-
рование

Фрагмент сети лингвистической онтологии (многозначные термины «театр», «балет» и др.)



Фрагмент описания предметной области





/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. (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.** (док./стр.)

☐ 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

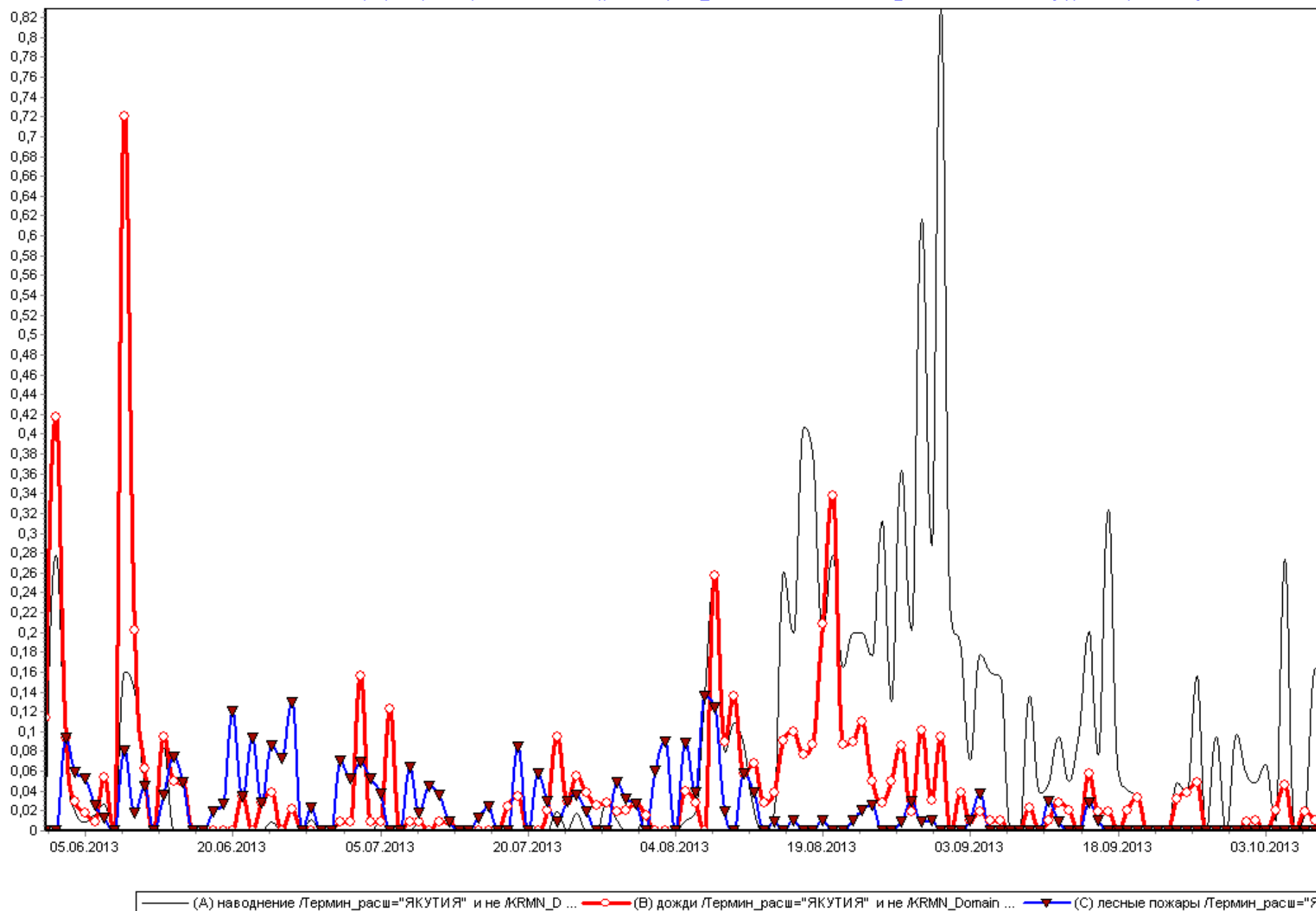
На повестке дня вопрос о **помощи** пострадавшим в результате паводка на Дальнем Востоке. Внимание высшего руководства страны к людям, оказавшимся в беде из-за стихии, вполне объяснимо: власть должна выполнить свою работу и защитить граждан.

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

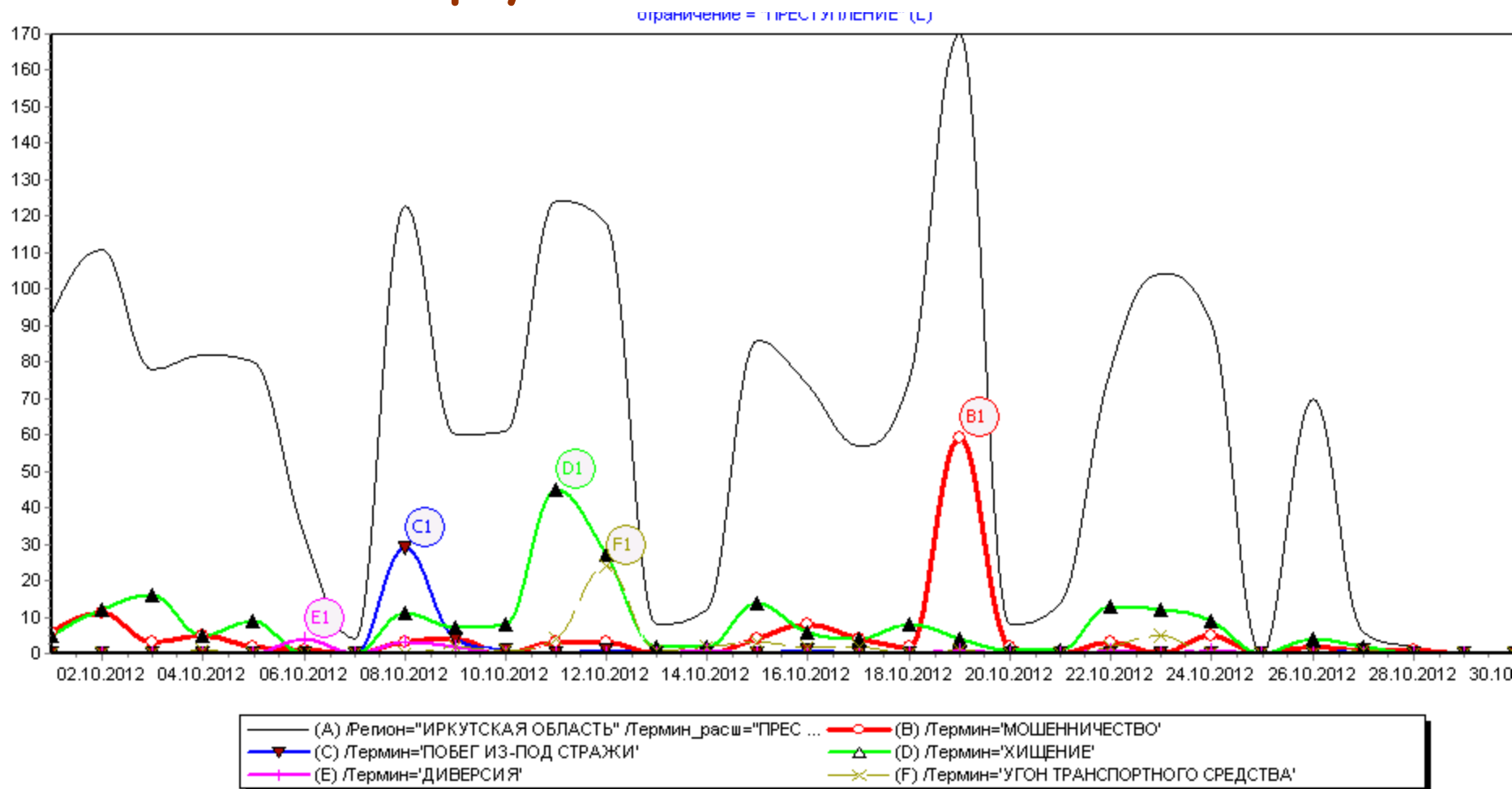
+/-		Термин
+	+t	ПРИАМУРЬЕ
-	-t	
+	+t	БЛАГОВЕЩЕНСК
-	-t	
+	+t	АМУРСКАЯ ОБЛАСТЬ
-	-t	
+	+t	СОУЧАСТИЕ
-	-t	
+	+t	РАСЧЕТНЫЙ СЧЕТ
-	-t	
+	+t	ЕВРЕЙСКАЯ АВТОНОМНАЯ ОБЛАСТЬ
-	-t	
+	+t	АМУР
-	-t	

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]

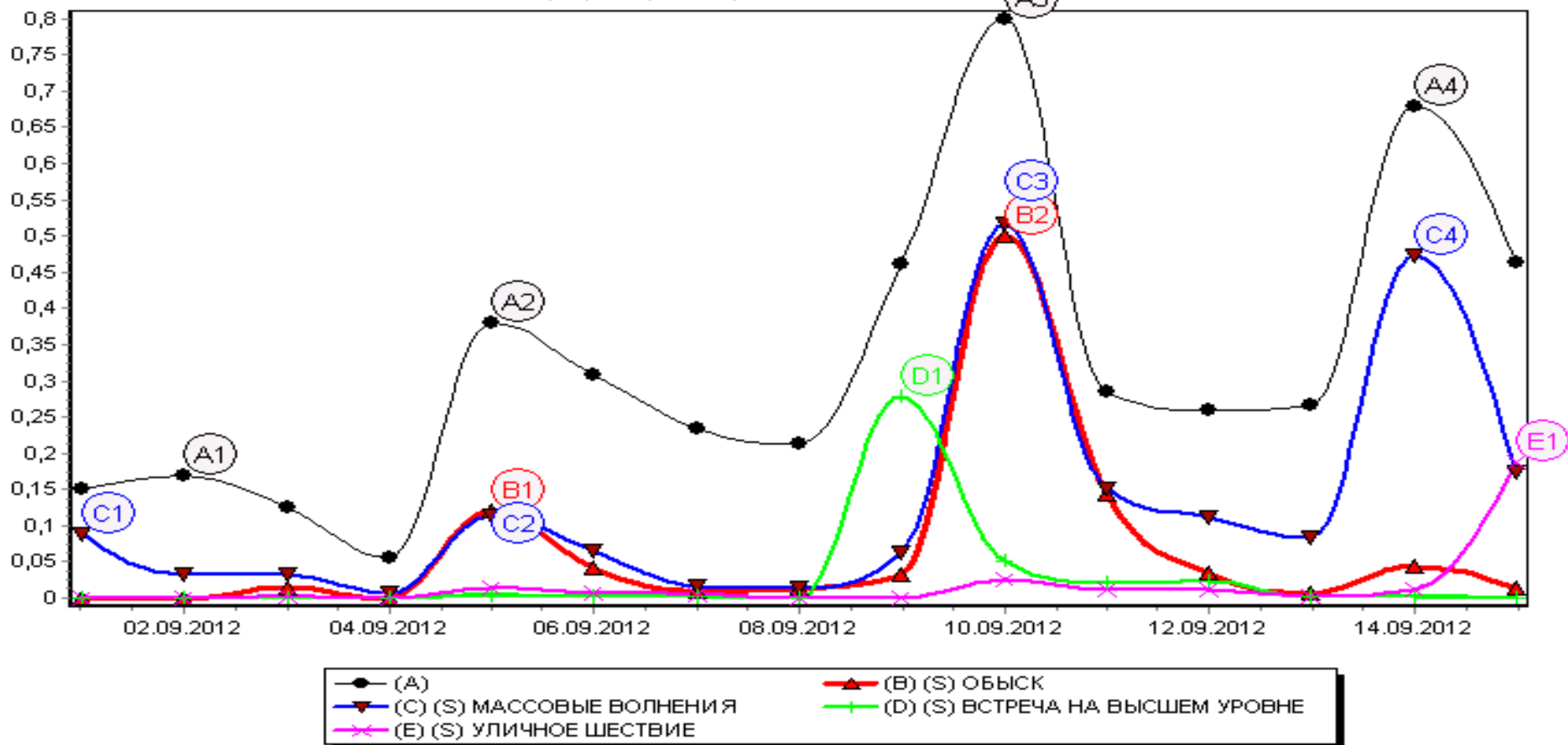


Спектрально-фасетный анализ преступлений в Иркутской области по видам



B1 19.10.2012 В Москве по подозрению в покушении на мошенничество задержан экс-чиновник из Иркутской области //gubmvd.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 №6 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

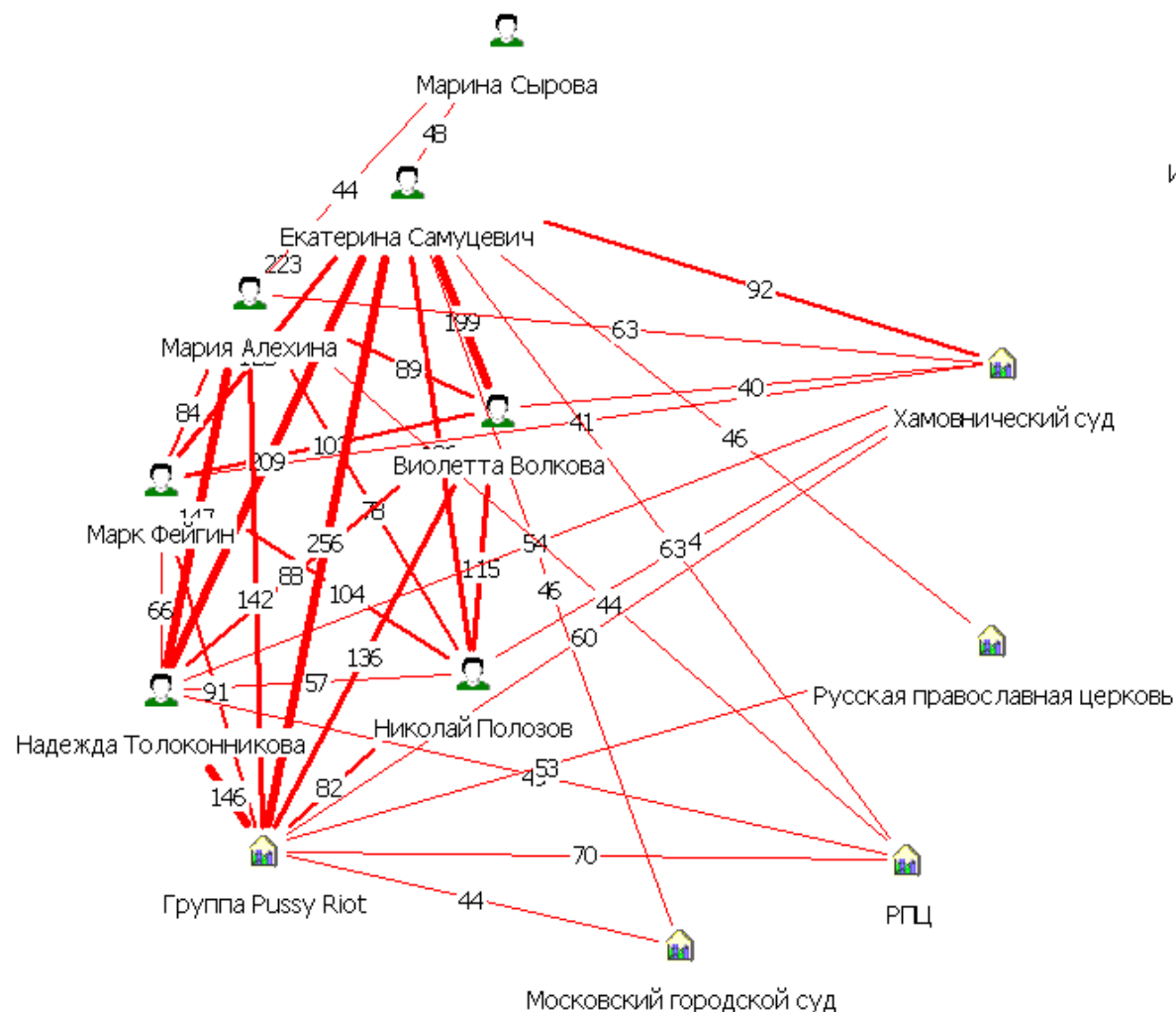
D1 09.09.2012 16:05:00 09.09.2012 16 05 Ксения Собчак была немало удивлена тем, что запись в ее твиттере обсуждал президент Владим

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)



Добавление в когнитивную схему иерархий лингвистической онтологии

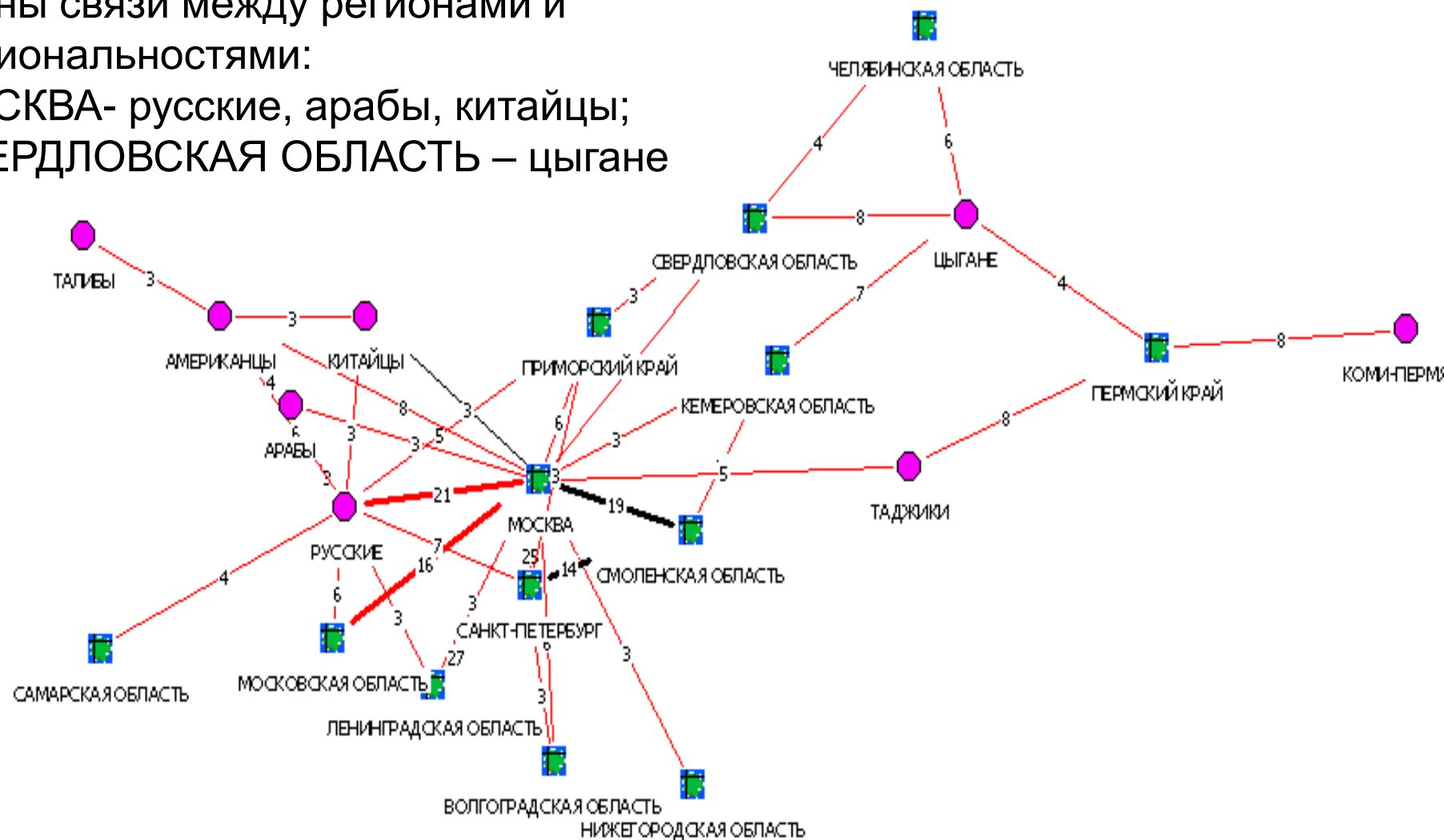
Запрос: «героин /Дата_док="09.2012"»

видны связи между регионами и

национальностями:

МОСКВА- русские, арабы, китайцы;

СВЕРДЛОВСКАЯ ОБЛАСТЬ – цыгане



Построение аналитических отчетов

- **Задание темы исследования**
- **Формирование условий запросов**
- **Формирование правил структурирования**
- **Поиск релевантных фрагментов**
- **Кластеризация фрагментов в заданной структуре**
- **Формирование отчетного документа**
- **Редактирование отчетного документа**
- **Подверстывание карт, графиков, семантических схем**
- **Вывод на печать**

Заголовок (запрос)

подрубрика₁

фрагмент₁₁

фрагмент₁₂

фрагмент₁₃

подрубрика₂

фрагмент₂₁

фрагмент₂₂

фрагмент₂₃

фрагмент₂₄

подрубрика₃

фрагмент₃₁

фрагмент₃₂

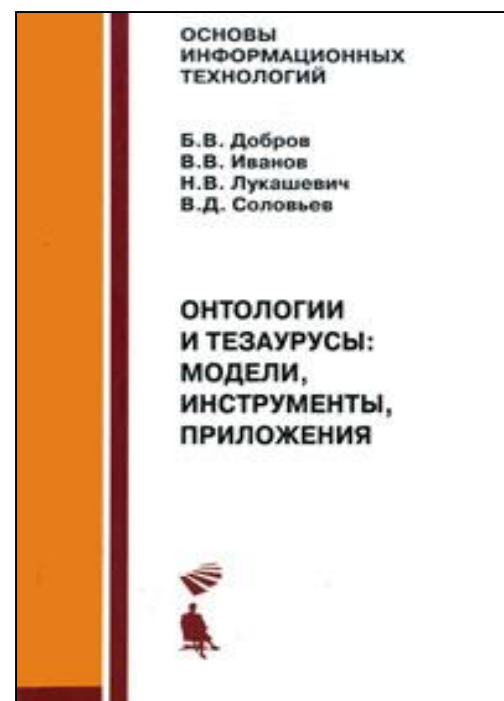
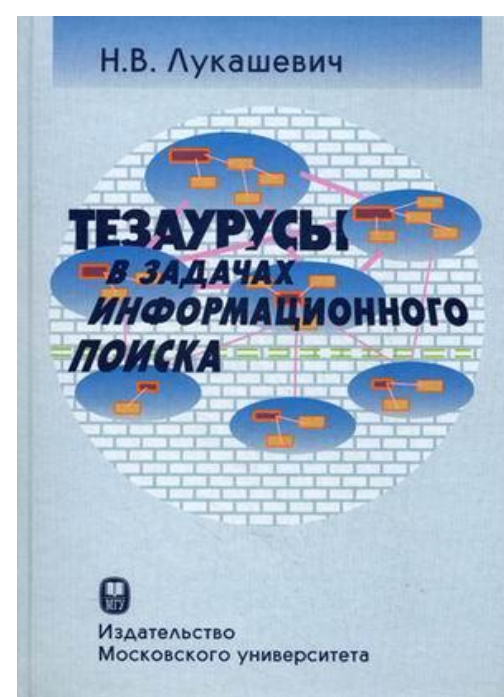
Основные проекты

[illegible]

Всего более 140 публикаций

Некоторые последние:

- Chetviorkin I.I., Loukachevitch N.V., Sentiment analysis track at ROMIP 2012 // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 г.). Вып. 12 (19). — М.: Изд-во РГГУ, 2013
- Michael Nokel, Elena Bolshakova, Natalia Loukachevitch, Topic Models Can Improve Domain Term Extraction // In Proceedings of the 35th European Conference on Information Retrieval (ECIR 2013)
- Ageev Mlikhail, Lagun Dmitry, Emory University, Agichtein Eugene, Improving Search Result Summaries By Using Searcher Behavior Data // In Proceedings of the 36th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '13)
- Iliia Chetviorkin, Natalia Loukachevitch, Extraction of Russian Sentiment Lexicon for Product Meta-Domain // In Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)
- Mikhail Ageev, Qi Guo, Dmitry Lagun, and Eugene Agichtein. Find it if you can: a game for modeling different types of web search success using interaction data // In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '11)
- Dobrov B., Loukachevitch N. Multiple Evidence for Term Extraction in Broad Domains // In Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP-2011)
- Loukachevitch Natalia. Establishment of taxonomic relationships in linguistic ontologies. Lecture Notes in Computer Science, 2011, V.6581, Knowledge processing and data analysis.



Заключение

- Расширяющийся круг прикладных задач
- Рассмотренные приложения: осязаемые результаты
- В основном используются простые и редуцированные модели ЕЯ – причина: трудоемкость разработки сложных моделей, неэффективность соответствующих алгоритмов
- Современная тенденция – применение машинного обучения, которое дополняет традиционный подход
- *Rule-based* подход (инженерный) - основан на правилах, имеющих лингвистическую интерпретацию

Основная литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011.
2. Manning Ch., Schutze H. Foundations of Statistical Natural Language Processing.
3. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. Вильямс, 2011. <http://nlp.stanford.edu/IR-book/>
4. Сегалович И. Как работают поисковые системы — . «Мир Интернет», №10. - 2002 .
5. Пескова О. В. Методы автоматической классификации электронных текстовых документов без обучения [кластеризация текстов] // Научно-техническая информация. Сер. 2. – 2006. – № 12. – С. 21-32.
6. Кобозева И.М. Лингвистическая семантика. – М., 2009.
7. Jurafsky D., Martin J. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Prentice Hall, 2000

Доп. литература

1. Васильев В. Г., Кривенко М. П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008.
2. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во Московского университета, 2011.
3. Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.