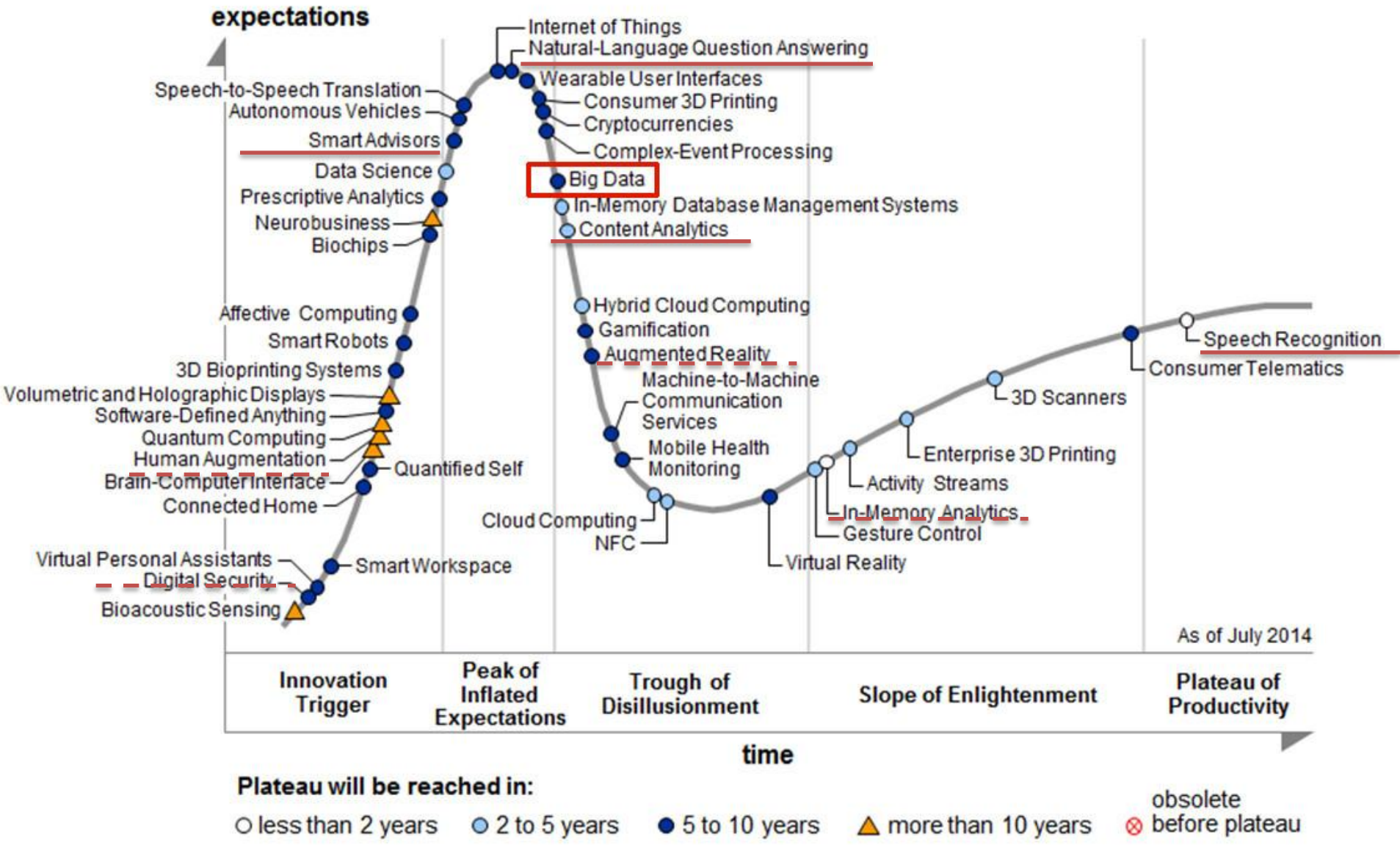
Данные в текстовой аналитике

Борис Добров, Наталья Лукашевич
Научно-исследовательский вычислительный центр
МГУ имени М.В. Ломоносова

Data in Text Analytics

Boris Dobrov, Natalia Loukachevitch
Research Computing Center of
M.V. Lomonosov Moscow State University

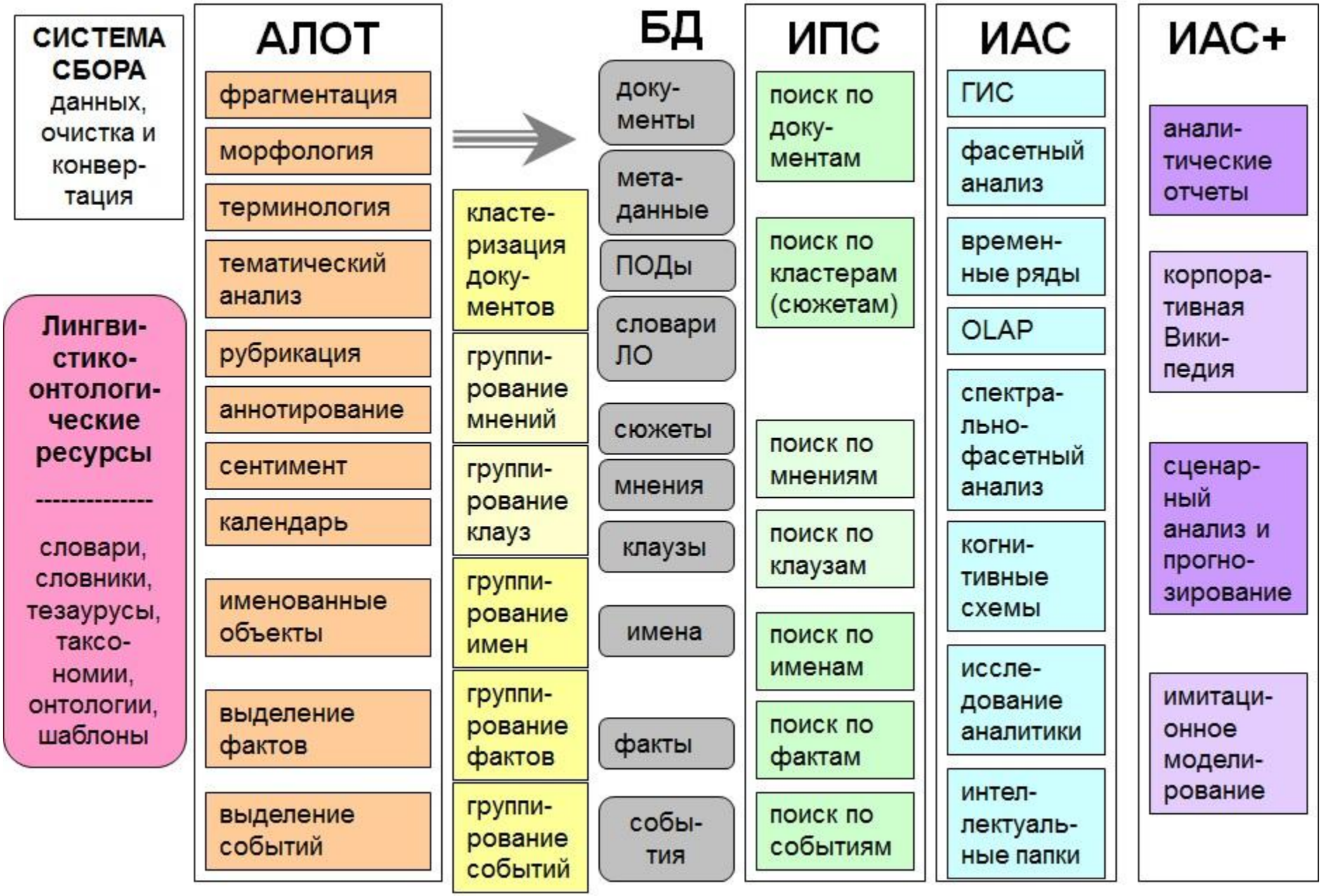
Gartner Curve of Advanced Technologies



Enterprise Level vs. Global Services
Natural Language Processing /Information Retrieval

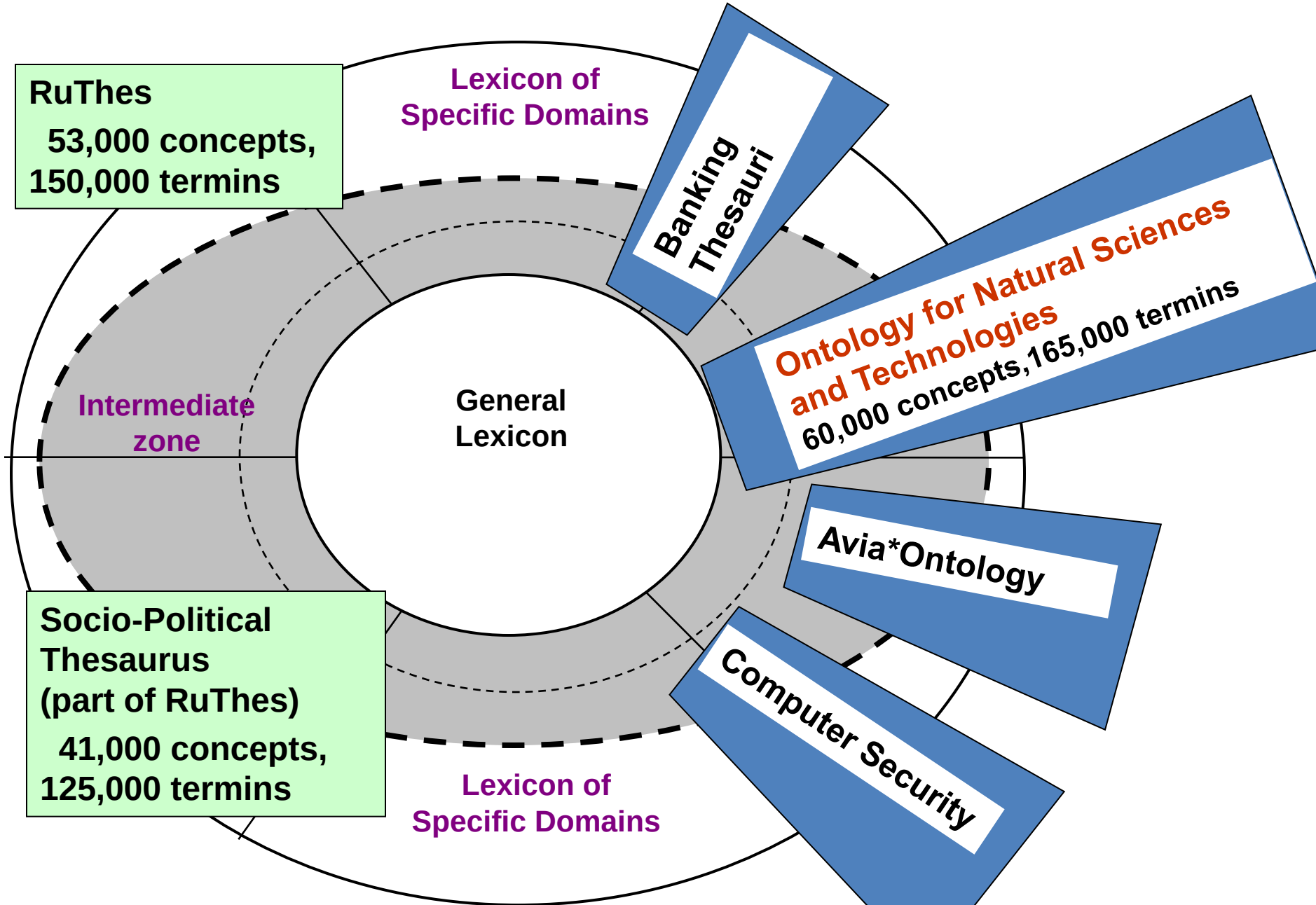
- High requirements to the quality
  (errors on the main portal page is impossible,
  manual post-editing)
- Not the user's fault, only the software
- Enough large corpora
  (news for the month ~ 1 million documents,
  integrally ~ 100 million documents)
- A special business role = analytist:
  - briefly describe what is happening
  - try to explain (main active actors, forces, causes)
  - if possible, try to predict

Task Flow

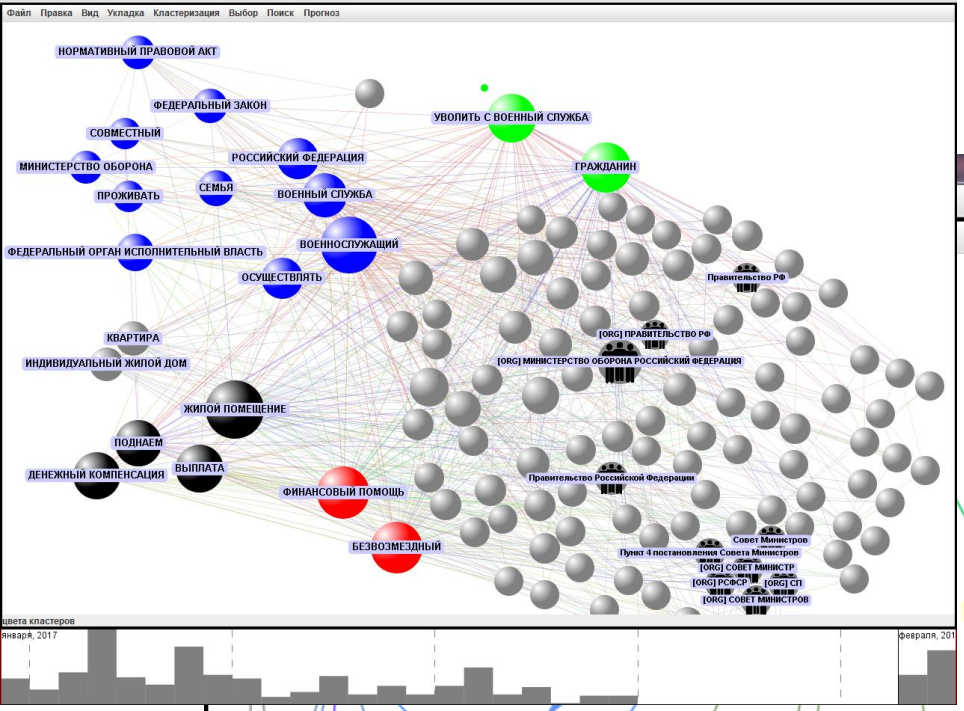

СИСТЕМА СБОРА данных, очистка и конвер-тация
Лингви-стико-онтологи-ческие ресурсы

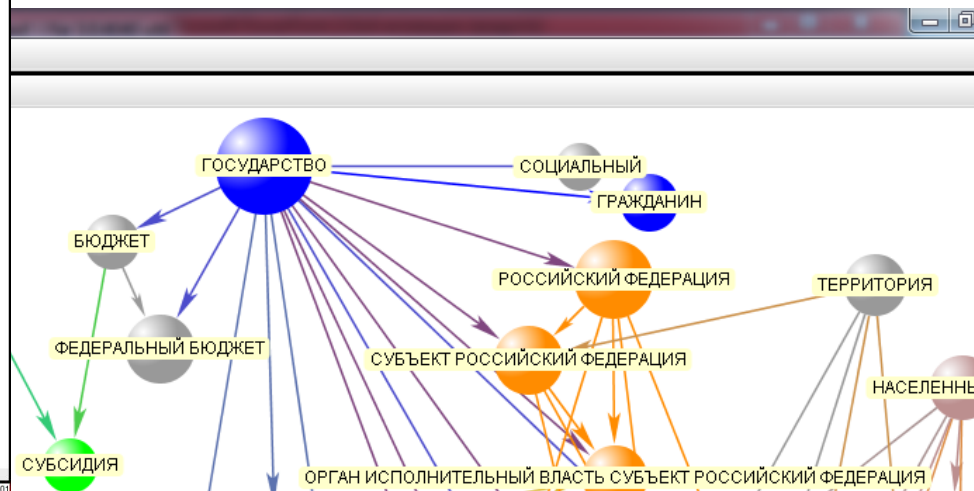
словари, словники, тезаурусы, табсо-номии, онтологии, шаблоны
АЛОТ
фрагментация
морфология
терминология
тематический анализ
рубрикация
аннотирование
сентимент
календарь
именованные объекты
выделение фактов
выделение событий
кластеризация документов
группирование мнений
группирование клауз
группирование имен
группирование фактов
группирование событий
БД
документы
метаданные
ПОДы
словари ЛО
сюжеты
мнения
клаузы
имена
факты
события
ИПС
поиск по документам
поиск по кластерам (сюжетам)
поиск по мнениям
поиск по клаузам
поиск по именам
поиск по фактам
поиск по событиям
ИАС
ГИС
фасетный анализ
временные ряды
OLAP
спектрально-фасетный анализ
когнитивные схемы
исследование аналитики
интеллектуальные папки
ИАС+
аналитические отчеты
корпоративная Википедия
сценарный анализ и прогнозирование
имитационное моделирование

Large Linguistics Ontologies

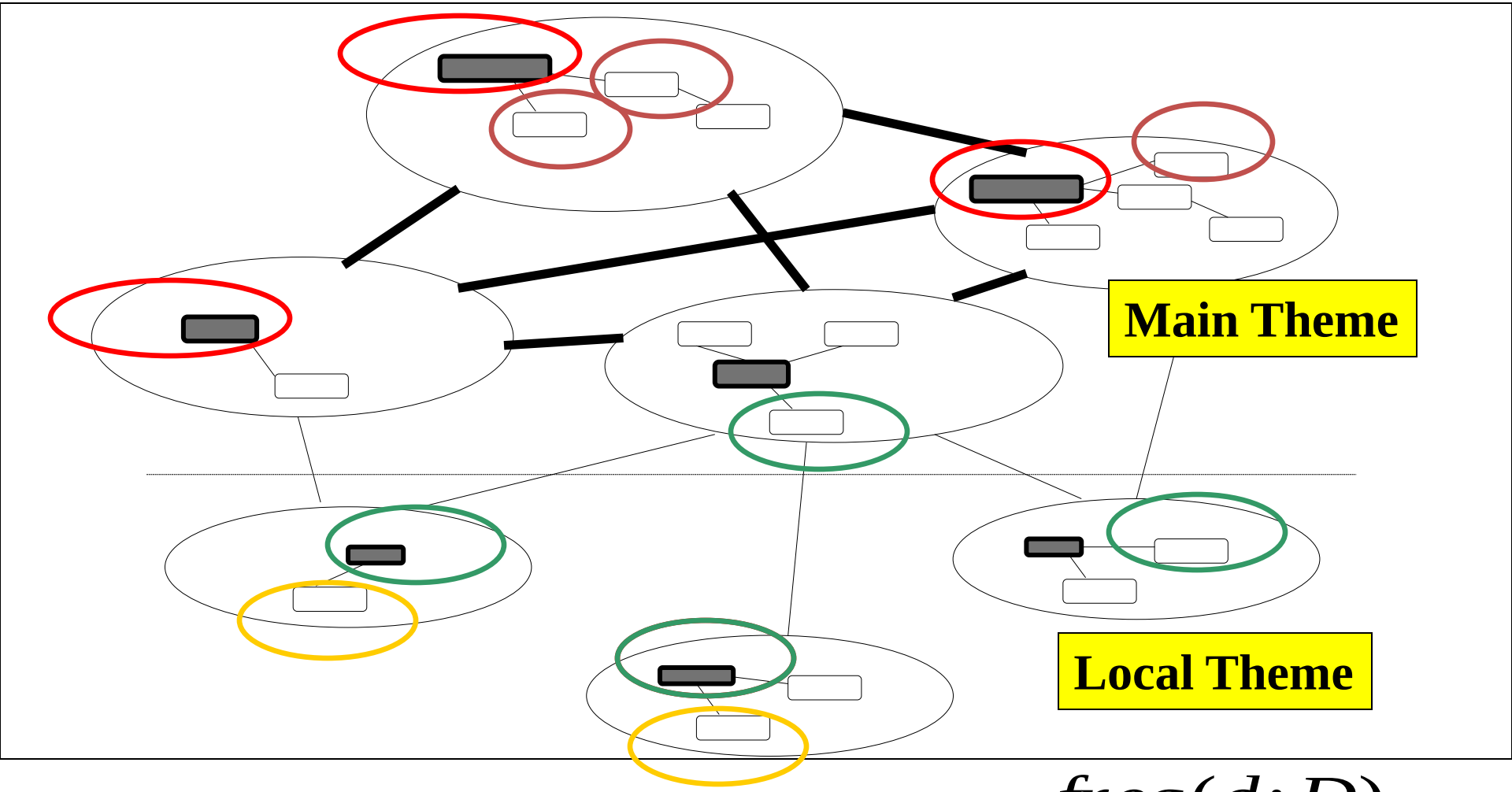


Представления текста





Thematic Representation



$$\theta(d) = \alpha \cdot \omega(d;D) + (1-\alpha) \cdot \frac{freq(d;D)}{\max_{c} freq(c;D)}$$

▼ ☑ ОБРАБОТАННЫЙ ТЕКСТ

Постановление Правительства РФ от 26 июня 1995 г. N 604

"О **порядке** оказания **безвозмездной финансовой помощи** на **строительство** (**покупку) жилья** и **выплаты денежной компенсации** за **наем** (**поднаем**) **жилых помещений военнослужащим** и **гражданам**, **уволенным с военной службы**"

Во **исполнение Закона Российской Федерации** "О **статусе военнослужащих**" и в целях **обеспечения прав на жилище военнослужащих** и **граждан**, **уволенных с военной службы**, **Правительство Российской Федерации постановляет**:

1. **Утвердить** прилагаемое **Положение о порядке** оказания **безвозмездной финансовой помощи на строительство** (**покупку) жилья** и **выплаты денежной компенсации** за **наем** (**поднаем**) **жилых помещений военнослужащим** и **гражданам**, **уволенным с**

2. **Министерству обороны Российской Федер**
которых предусмотрена **военная служба**:

в месячный **срок разработать** и **утвердить** фо
решения об оказании военнослужащим безв
жилья и о **выплате денежной компенсации** за

расходы, связанные **с оказанием военнослуж**
денежной компенсации за **наем** (**поднаем**) **жи**
предусматриваемых **федеральным органам ис**
цели.

3. **Органам исполнительной власти субъекто**

оказывать безвозмездную финансовую пом
уволенным с военной службы, **осуществляю**
средств федерального бюджета, **выделяемы**
Информация об **изменениях**:

выплачивать денежную компенсацию гражд
помещений при невозможности **обеспечения ж**
срок со дня их **прибытия** к избранному **постоя**
предусматриваемых в **федеральном бюджете**

предусматривать при формировании бюджета

☑ ТЕМАТИЧЕСКАЯ АННОТАЦИЯ

****							ВОЕННОСЛУЖАЩИЙ; ВОЕННОСЛУЖАЩИЙ-КОНТРАКТНИК; ВОЕННАЯ СЛУЖБА; УВОЛЬНЕНИЕ С ВОЕННОЙ СЛУЖБЫ;
****	●						ФИНАНСОВАЯ ПОМОЩЬ; СУБСИДИЯ;
****	●	●					ГРАЖДАНИН; ГОСУДАРСТВО;
****	●	●	●				ФИНАНСОВЫЕ РАСЧЕТЫ; ДЕНЕЖНОЕ ДОВОЛЬСТВИЕ; ДЕНЕЖНАЯ КОМПЕНСАЦИЯ; БЕЗНАЛИЧНЫЙ РАСЧЕТ;
****	●			●			НАЕМ ЖИЛОГО ПОМЕЩЕНИЯ;
****	●						РОССИЙСКАЯ ФЕДЕРАЦИЯ; ПРАВИТЕЛЬСТВО РСФСР; САНКТ-ПЕТЕРБУРГ; МОСКВА; СУБЪЕКТ РОССИЙСКОЙ ФЕДЕРАЦИИ; ОРГАН ИСПОЛНИТЕЛЬНОЙ ВЛАСТИ СУБЪЕКТА ФЕДЕРАЦИИ;
****	●	●	●	●	●		ПОКУПКА ЖИЛЬЯ; ПОКУПКА (ДЕЯТЕЛЬНОСТЬ); ОТДАТЬ, ДАТЬ, ПРЕДОСТАВИТЬ; ЖИЛОЕ ПОМЕЩЕНИЕ;

■ АННОТАЦИЯ

▼ ☑ ОБРАБОТАННЫЙ ТЕКСТ

Постановление Правительства РФ от 26 июня 1995 г. N 604

"О **порядке** оказания **безвозмездной финансовой помощи на строительство** (**покупку) жилья** и **выплаты денежной компенсации** за **наем** (**поднаем**) **жилых помещений военнослужащим** и **гражданам**, **уволенным с военной службы**"

Во **исполнение Закона Российской Федерации** "О **статусе военнослужащих**" и в целях **обеспечения прав на жилище военнослужащих** и **граждан**, **уволенных с военной службы**, **Правительство Российской Федерации постановляет**:

1. **Утвердить** прилагаемое **Положение о порядке** оказания **безвозмездной финансовой помощи на строительство** (**покупку) жилья** и **выплаты денежной компенсации** за **наем** (**поднаем**) **жилых помещений военнослужащим** и **гражданам**, **уволенным с военной службы**.

2. **Министерству обороны Российской Федерации** и иным **федеральным органам исполнительной власти**, в которых предусмотрена **военная служба**:

в месячный **срок разработать** и **утвердить** формы и **перечень документов**, необходимых **для принятия решения об оказании военнослужащим безвозмездной финансовой помощи на строительство** (**покупку) жилья** и о **выплате денежной компенсации** за **наем** (**поднаем**) **жилых помещений**;

расходы, связанные **с оказанием военнослужащим безвозмездной финансовой помощи** и **выплатой денежной компенсации** за **наем** (**поднаем**) **жилых помещений**, **производить** за **счет** и в пределах **средств**, предусматриваемых **федеральным органам исполнительной власти** в **федеральном бюджете** на указанные цели.

☑ ТЕМАТИЧЕСКАЯ АННОТАЦИЯ

****							**ВОЕННОСЛУЖАЩИЙ; ВОЕННОСЛУЖАЩИЙ-КОНТРАКТНИК; ВОЕННАЯ СЛУЖБА; УВОЛЬНЕНИЕ С ВОЕННОЙ СЛУЖБЫ;**
****	●						**ФИНАНСОВАЯ ПОМОЩЬ; СУБСИДИЯ;**
****	●	●					**ГРАЖДАНИН; ГОСУДАРСТВО;**
****	●	●	●				**ФИНАНСОВЫЕ РАСЧЕТЫ; ДЕНЕЖНОЕ ДОВОЛЬСТВИЕ; ДЕНЕЖНАЯ КОМПЕНСАЦИЯ; БЕЗНАЛИЧНЫЙ РАСЧЕТ;**
****	●			●			**НАЕМ ЖИЛОГО ПОМЕЩЕНИЯ;**
****	●						**РОССИЙСКАЯ ФЕДЕРАЦИЯ; ПРАВИТЕЛЬСТВО РСФСР; САНКТ-ПЕТЕРБУРГ; МОСКВА; СУБЪЕКТ РОССИЙСКОЙ ФЕДЕРАЦИИ; ОРГАН ИСПОЛНИТЕЛЬНОЙ ВЛАСТИ СУБЪЕКТА ФЕДЕРАЦИИ;**
****	●	●	●	●	●		**ПОКУПКА ЖИЛЬЯ; ПОКУПКА (ДЕЯТЕЛЬНОСТЬ); ОТДАТЬ, ДАТЬ, ПРЕДОСТАВИТЬ; ЖИЛОЕ ПОМЕЩЕНИЕ;**

■ АННОТАЦИЯ

▼ ☑ ОБРАБОТАННЫЙ ТЕКСТ

Постановление Правительства РФ от 26 июня 1995 г. N 604

"О **порядке** оказания **безвозмездной финансовой помощи на строительство** (**покупку) жилья** и **выплаты денежной компенсации** за **наем** (**поднаем**) **жилых помещений военнослужащим** и **гражданам**, **уволенным с военной службы**"

Во **исполнение Закона Российской Федерации** "О **статусе военнослужащих**" и в целях **обеспечения прав на жилище военнослужащих** и **граждан**, **уволенных с военной службы**, **Правительство Российской Федерации постановляет**:

1. **Утвердить** прилагаемое **Положение о порядке** оказания **безвозмездной финансовой помощи на строительство** (**покупку) жилья** и **выплаты денежной компенсации** за **наем** (**поднаем**) **жилых помещений военнослужащим** и **гражданам**, **уволенным с военной службы**.

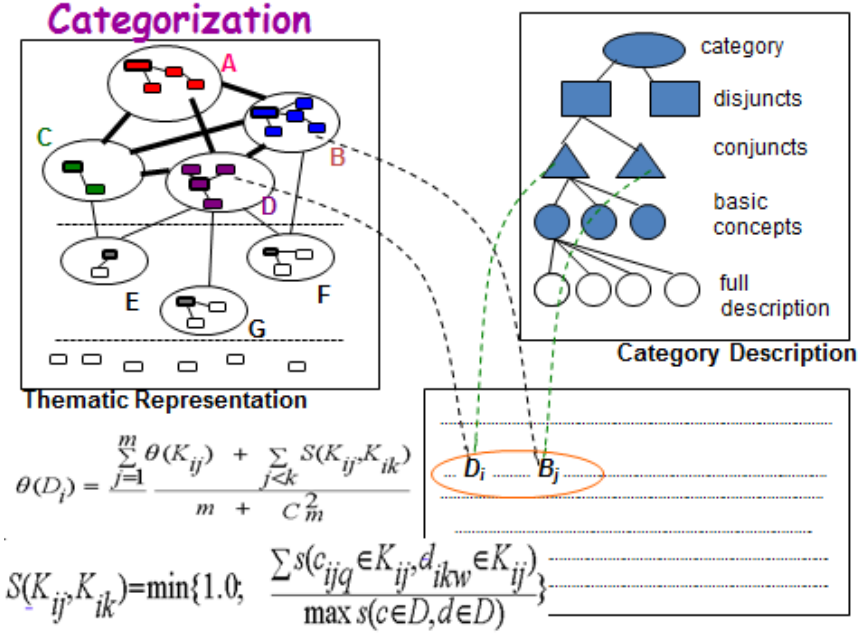
2. **Министерству обороны Российской Федерации** и иным **федеральным органам исполнительной власти**, в которых предусмотрена **военная служба**:

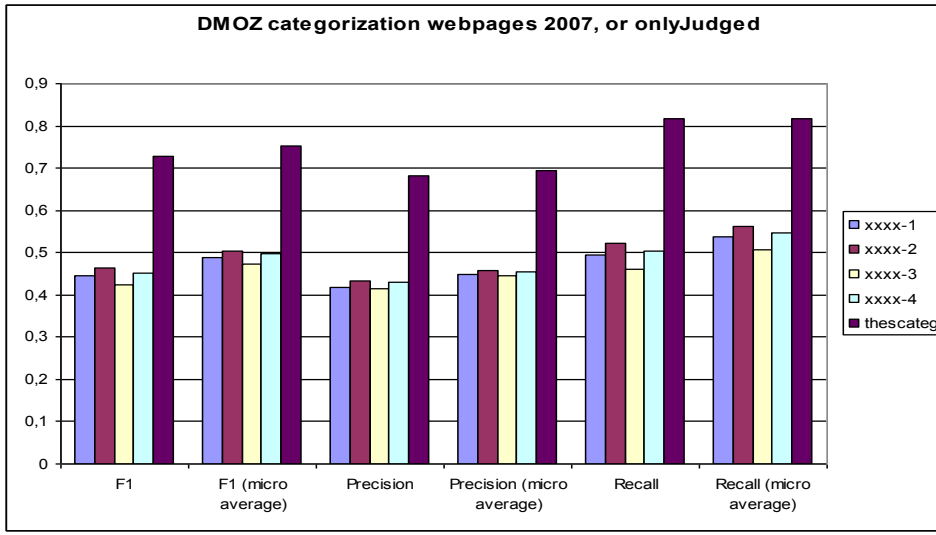
в месячный **срок разработать** и **утвердить** формы и **перечень документов**, необходимых **для принятия решения об оказании военнослужащим безвозмездной финансовой помощи на строительство** (**покупку**) **жилья** и о **выплате денежной компенсации** за **наем** (**поднаем**) **жилых помещений**;

расходы, связанные **с оказанием военнослужащим безвозмездной финансовой помощи** и **выплатой денежной компенсации** за **наем** (**поднаем**) **жилых помещений**, **производить** за **счет** и в пределах **средств**,

Unified Approach. Evaluation

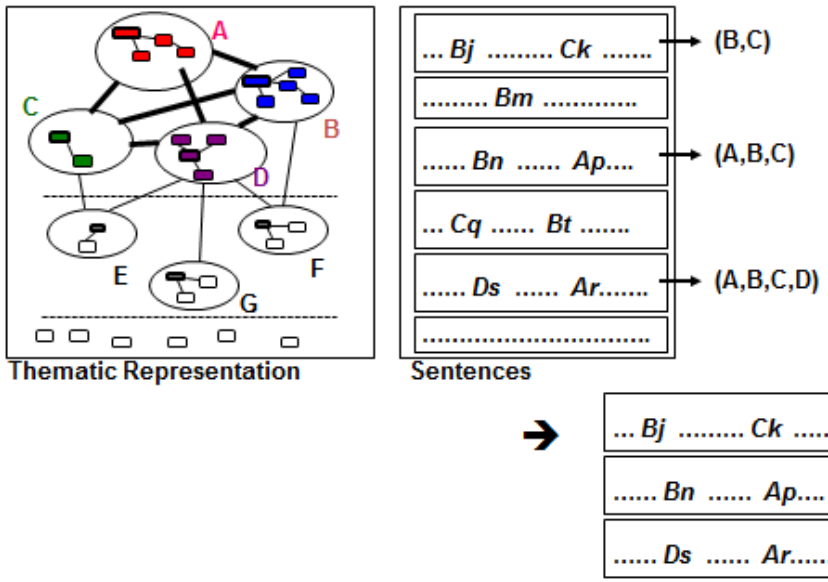
Categorization

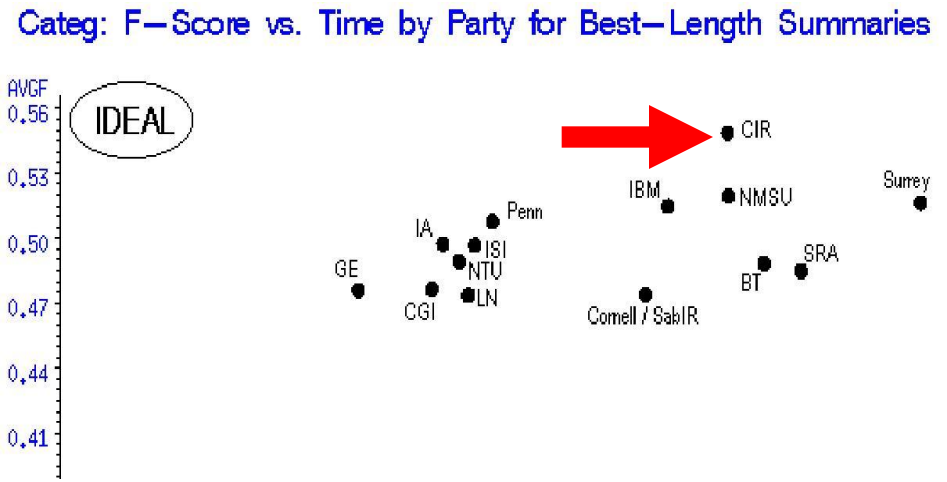




ROMIP2007 Web page categorization

Summarization





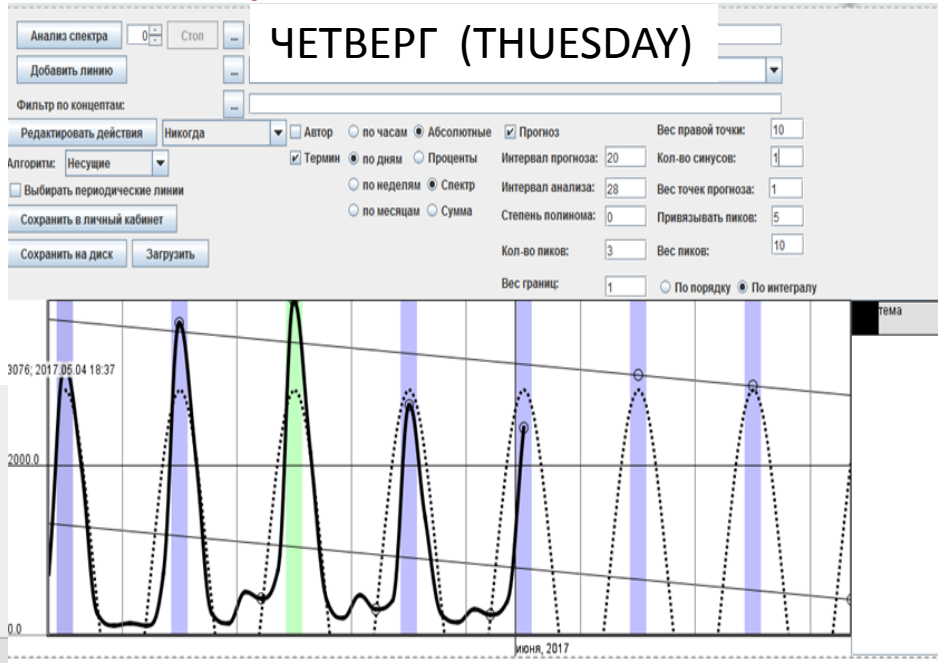
SUMMAC 1998 Summarization Evaluation (NIST DARPA TIPSTER III)

Основные проекты	Годы	ЛО: ОПТ	Новые ЛО	Поиск	QA	Рубрикация	Аннотирование	Кластеризация	Обзорное реферирование	Аналитические отчеты	Тональность
ГосДума ФС РФ	1999-н/в	✓		✓		✓	✓				
ЦБ РФ	2006-н/в	✓	✓			✓	✓		✓	✓	
Разные ведомства РФ	2000-н/в	✓	✓	✓		✓	✓	✓	✓	✓	✓
Яндекс	2014										✓
ГАС «Выборы» (ФКЗ «Право»)	1997-2011	✓	✓	✓	✓	✓					
НПП «Гарант-Сервис»	2002-2015	✓			✓	✓	✓	✓			
Рамблер. Новости	2008-2013	✓				✓	✓	✓	✓		
НП «Гидроэнергетика России»	2013	✓	✓	✓		✓	✓				
Минюст РФ	2007	✓		✓							
Счетная палата РФ	2003		✓								
ИППИ РАН (Упр. спецпрограмм)	1996	✓		✓		✓	✓				

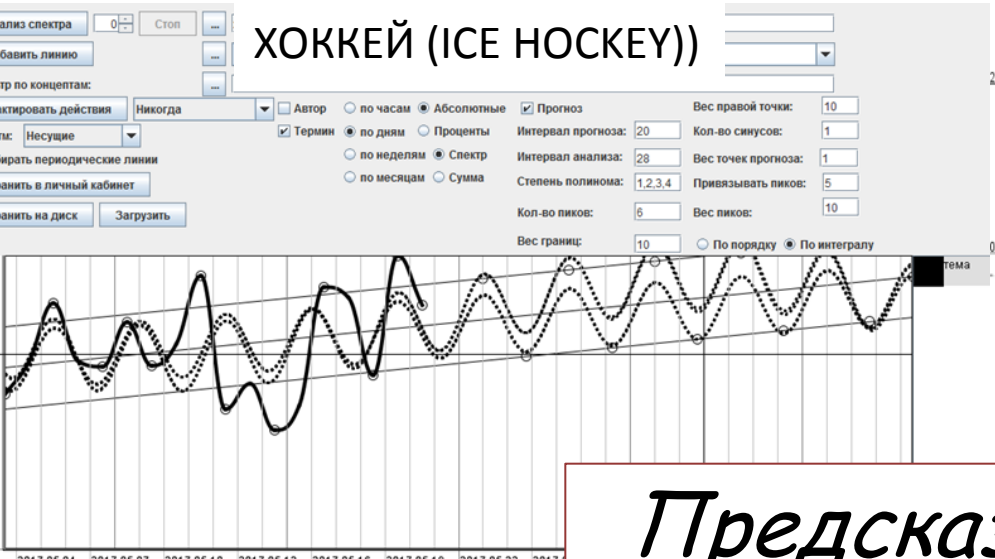
Main Tasks of NLP/ Information Retrieval

- Basic Natural Language Processing
  - Morphology — 99,9%
  - Named Entity Recognition
  - Categorization
  - Summarization — 60 % -75%
  - Sentiment
- Advances Natural Language Processing
  - Syntax
  - Keyword Extraction — 60 % -70%
  - Fact Extraction, Event Extraction
  - Clusterization
  - Multi-Document Summarization
- Desired Effect
  - Situation Awareness

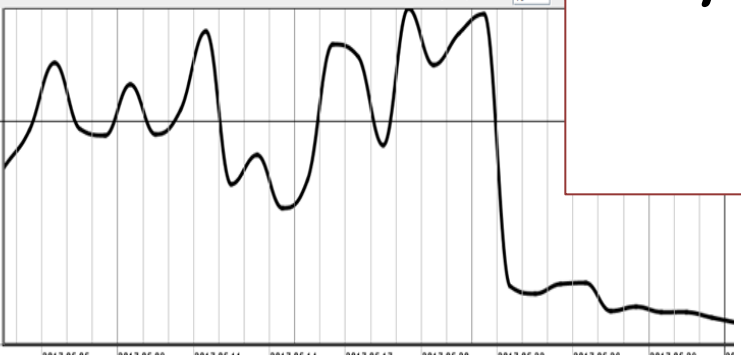
Time series analysis



ЧЕТВЕРГ (THUESDAY)



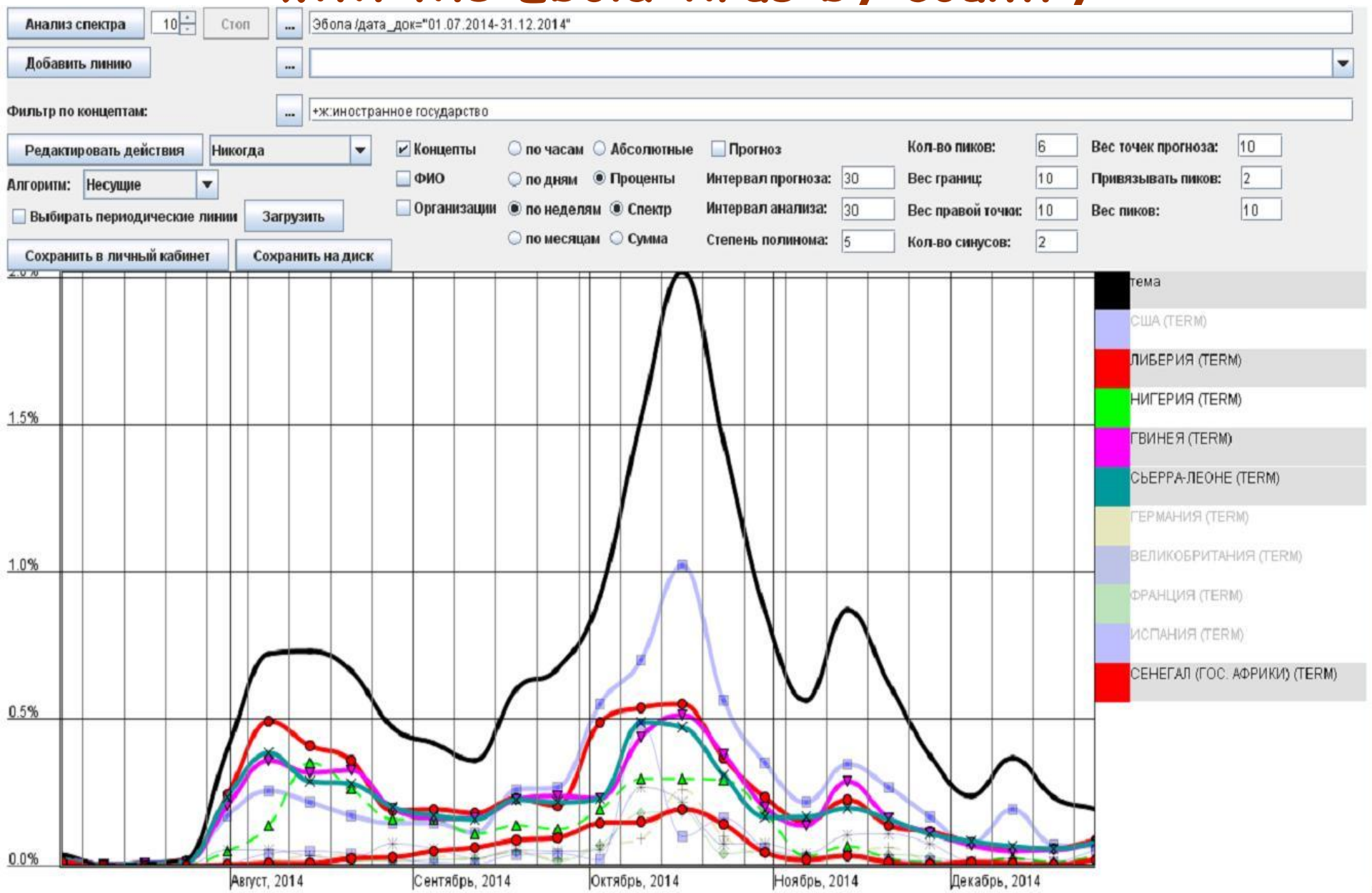
ХОККЕЙ (ICE HOCKEY))



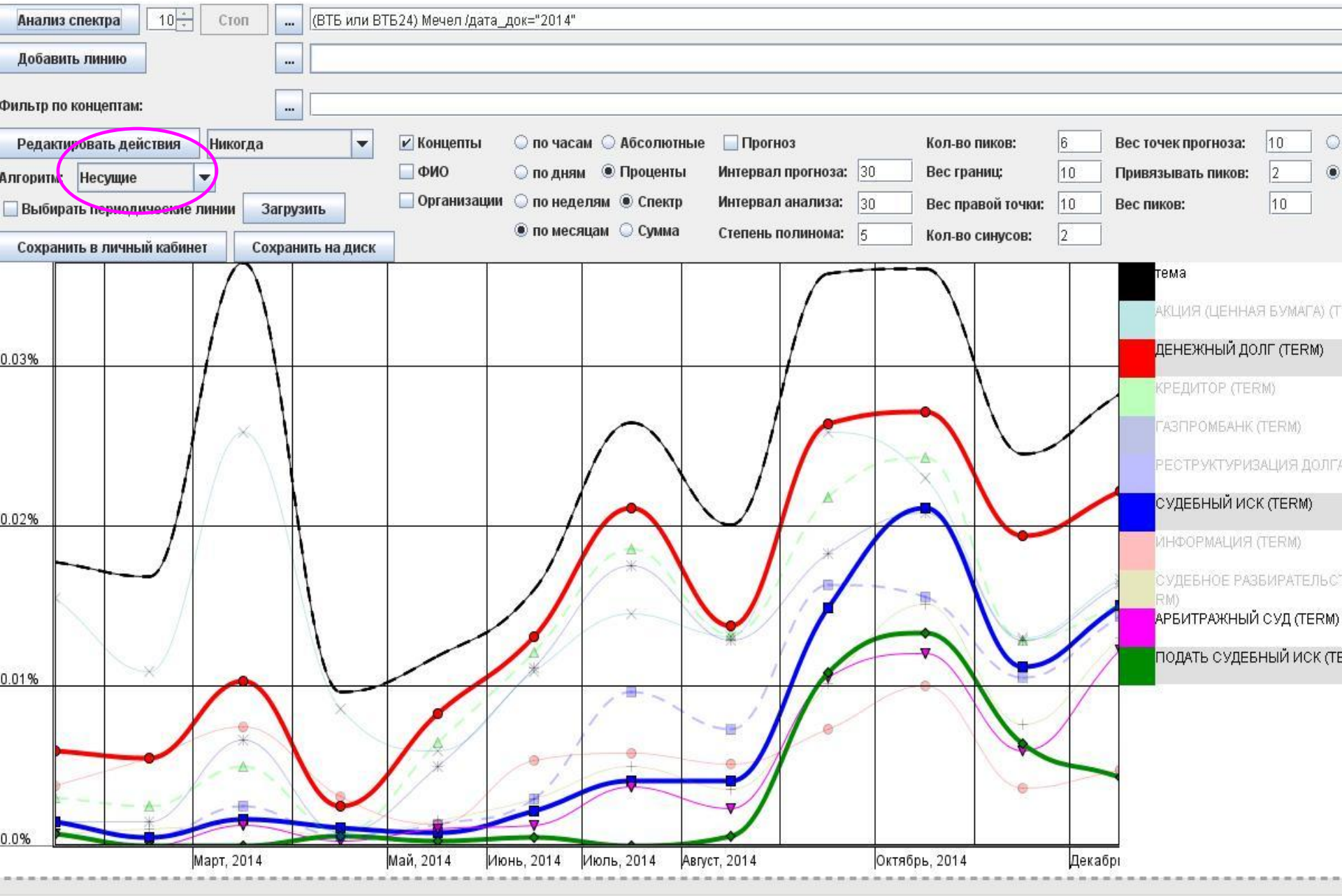
Предсказывать вообще трудно, особенно будущее

Виктор Степанович Черномырдин

The (simple) analysis of the situation with the Ebola virus by country



Time series analysis (Мечел vs. ВТБ)



noSQL Search Engine (semi-standard)

- Inverse Index

  [E_i [d_j, rank(E_i, d_j), [$positions_{ijk}$]]]

  -- search

  -- 1-2 sec in 1E+7 documents

- Direct Index

  [d_j [E_i , rank(E_i, d_j), [$positions_{ijk}$]]]

  -- analytics, highlight

  -- 0.3 sec for Top-200

  -- 5 sec for Top-2000

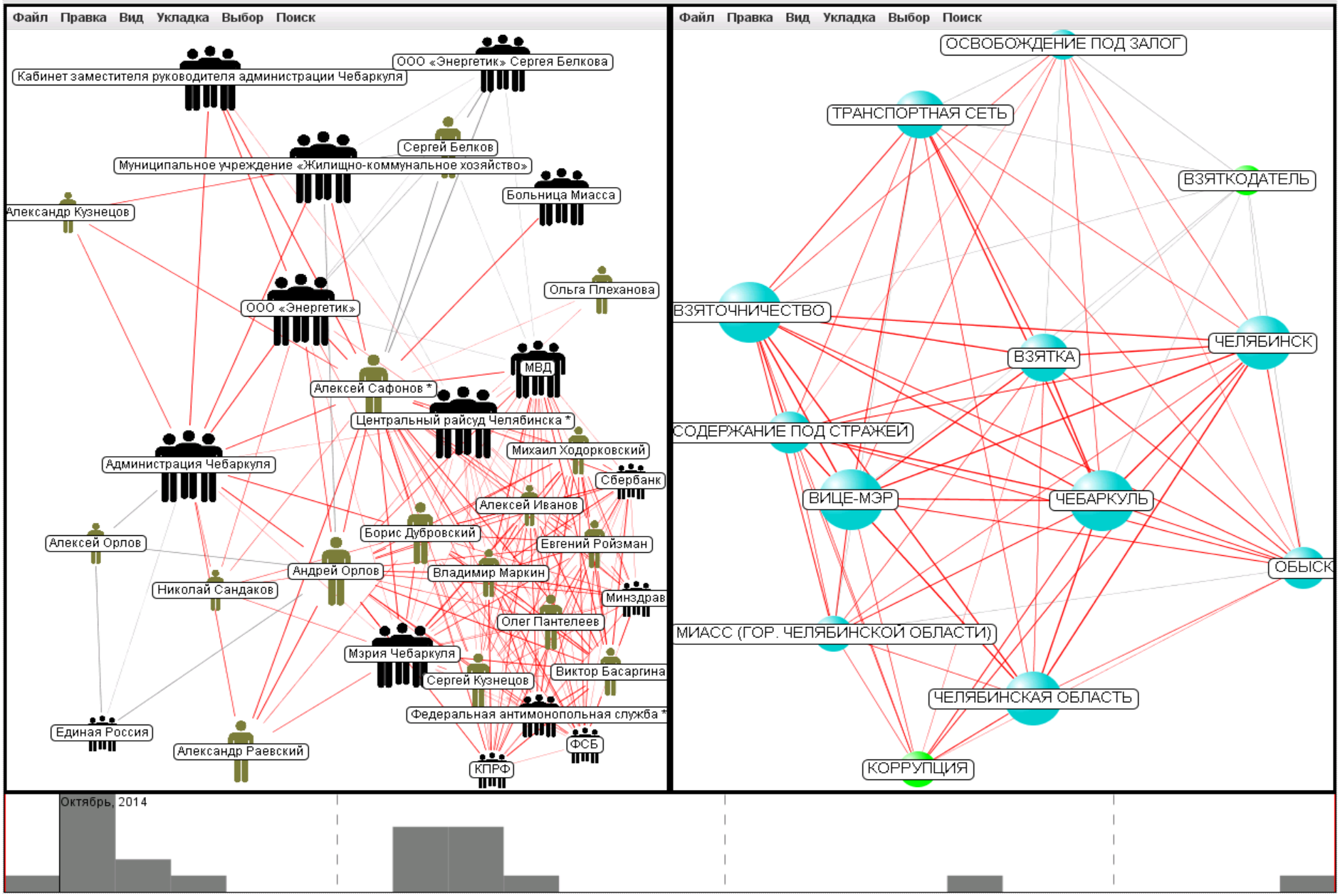
- Additional Indexes

  [E_i [$datetime_j$, doc_count(E_i , rank(E_i, d_j)>r_0)]]

  -- timeline analytics

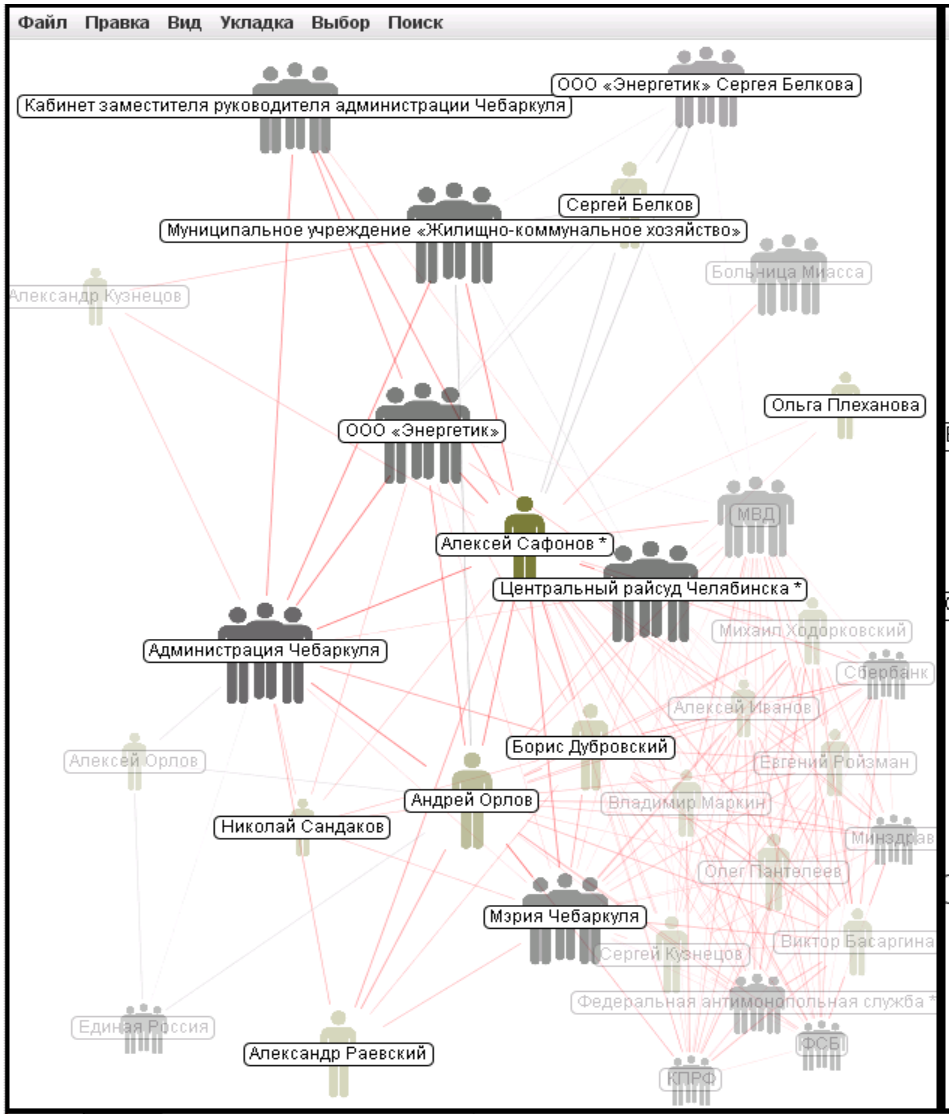
Cognitive graph

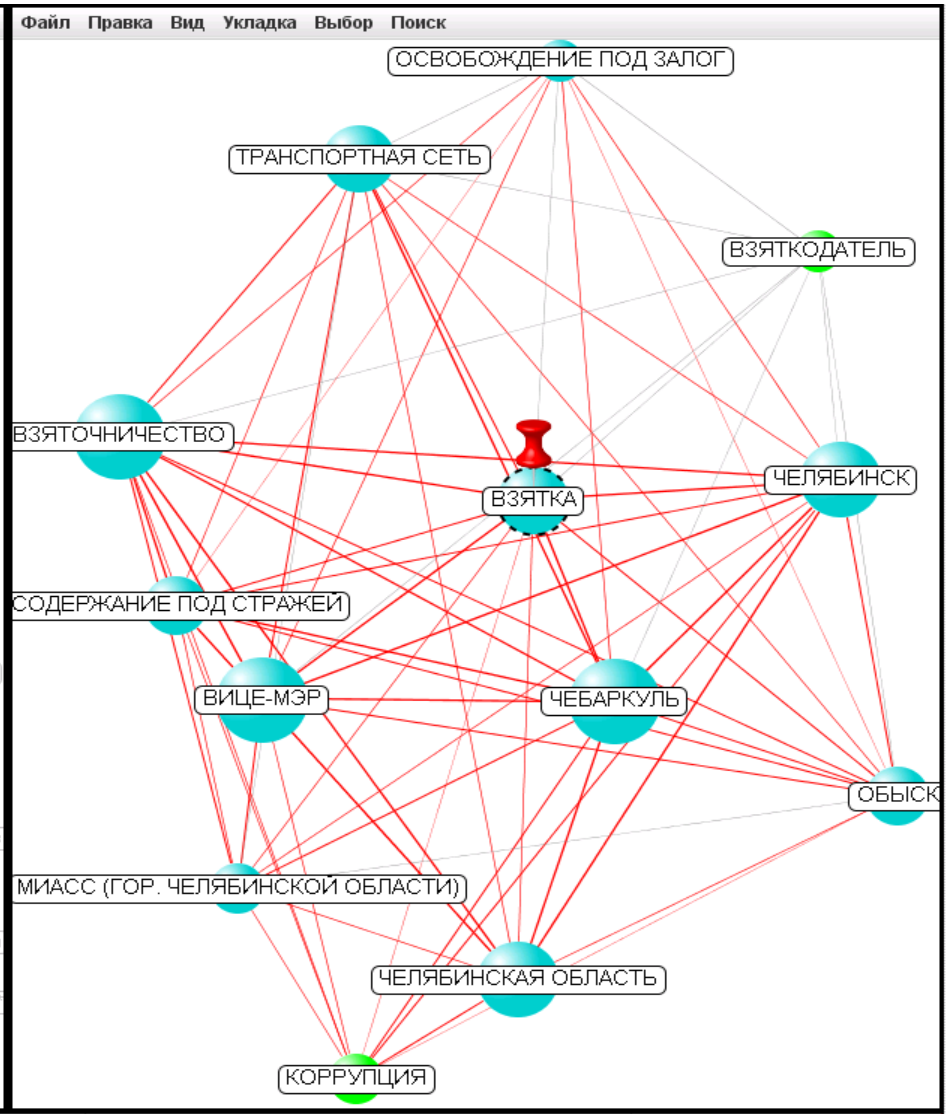
Query: Vice-major Chebarkul /date=«09.2014-10.2014»

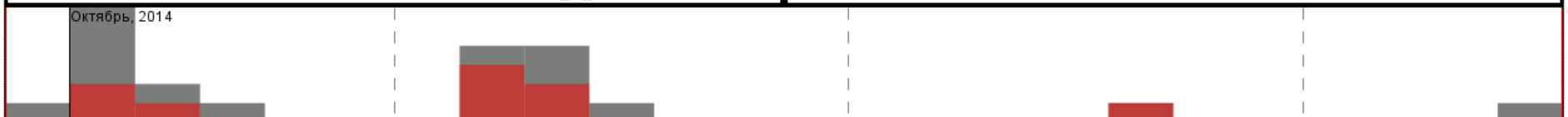


Dynamic analysis of a complex graph

Query: Vice-major Chebarkul /date=«09.2014-10.2014»

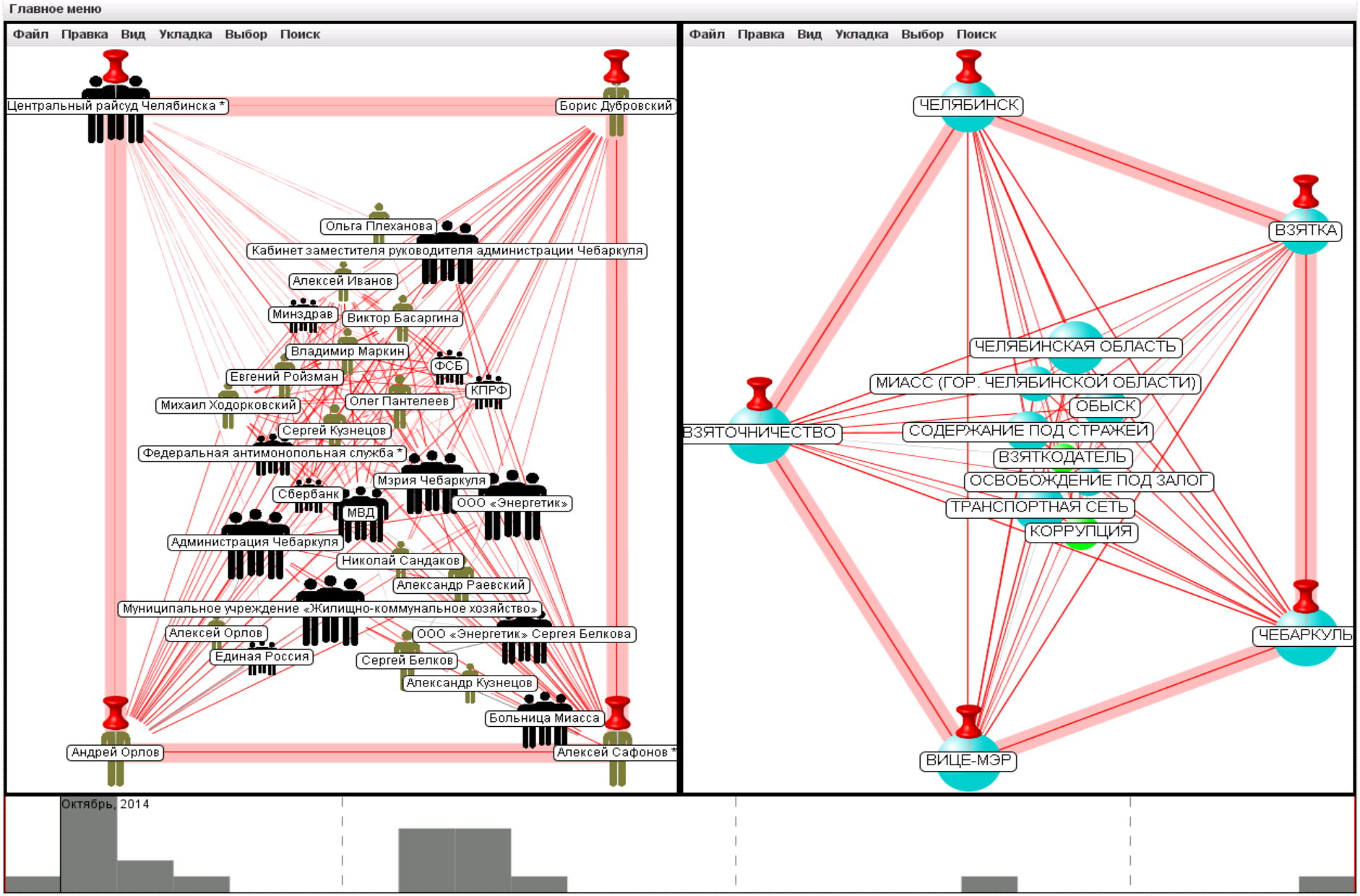






Main objects extraction

Query: Vice-major Chebarkul /date=«09.2014-10.2014»



Structured Analytic Report
Query: Vice-major Chebarkul (##conflicts)
Вице-мэр Чебаркуль (##конфликты)
(search+categorization+clusterization+MDS)

Классификатор: person.разбирательства
Сортировка: по возрастанию даты

00020 Конфликт

21.07.2014 18:14:00 **КАРТА КОНФЛИКТОВ УРФО: СЧЕТНАЯ ПАЛАТА ПРОТИВ КРСУ И МЭР ЧЕБАРКУЛЯ ПРОТИВ ДУБРОВСКОГО [FEDPRESS.RU]**

Карта конфликтов УрФО: Счетная палата против КРСУ и мэр Чебаркуля против Дубровского «УралПолит.Ru» подготовил очередной выпуск спецпроекта «Карта конфликтов УрФО». + 1 фрагм.
Вы узнаете, сколько миллиардов исчезли в «Корпорации развития Среднего Урала», почему мэр Чебаркуля оправдывается перед Дубровским, какие нарушения прокуратура нашла в скандальном югорском клубе и почему на Ямале растет число преступлений против детей. Познавательного чтения! Свердловская область Акции «Кольцово» ищут все. + 1 фрагм.
Поскольку реестр велся компанией, чей головной офис находится в Москве, была создана рабочая группа, написан соответствующий запрос. Вчера вечером мы получили выписку: 23 апреля была совершена передача акций на счет МУГИСО на основании передаточного письма от 23 апреля», – огорошил министр депутатов неожиданной новостью. Челябинская область Мэр Чебаркуля ответил штабу Дубровского на критику в свой адрес Стороны конфликта: мэр Чебаркуля Андрей Орлов, администрация Чебаркуля, Николай Сандаков, администрация Челябинской области, СМИ. + 1 фрагм.
Описание конфликта. Мэр Чебаркуля Андрей Орлов, попавший в озвученный на прошлой неделе «Рейтинг самых плохих глав муниципалитетов», ответил пиарщикам Дубровского, что он думает по поводу подобных списков. На своей страничке в Facebook глава администрации поделился не только мнением об антирейтингах, но и размышлениями о местном самоуправлении в России. + 1 фрагм.

29.07.2014 15:36:52 **ДУБРОВСКИЙ НАЗНАЧИЛ СВОЕГО ПРЕДСТАВИТЕЛЯ В ЧЕБАРКУЛЕ [URALINFORM.RU]**

Как сообщили "Уралинформбюро" в пресс-службе губернатора Южного Урала, 29 июля 2014 года Александр Раевский, бывший замминистра сельского хозяйства региона, назначен представителем главы области в Чебаркульском городском округе. "Приоритетные направления деятельности представителя - организация устойчивого функционирования объектов ЖКХ городского округа и контроль за целевым Вице-мэр Чебаркуля Сафонов требует от журналистов миллион рублей Вице-мэр Чебаркуля Сафонов требует от журналистов миллион рублей Сегодня в 13:56, просмотров: 67 Исполняющий обязанности заместителя главы Чебаркуля (Челябинская область) Алексей Сафонов подал в суд на информационное агентство «Ura.ru».

00140 Коррупция

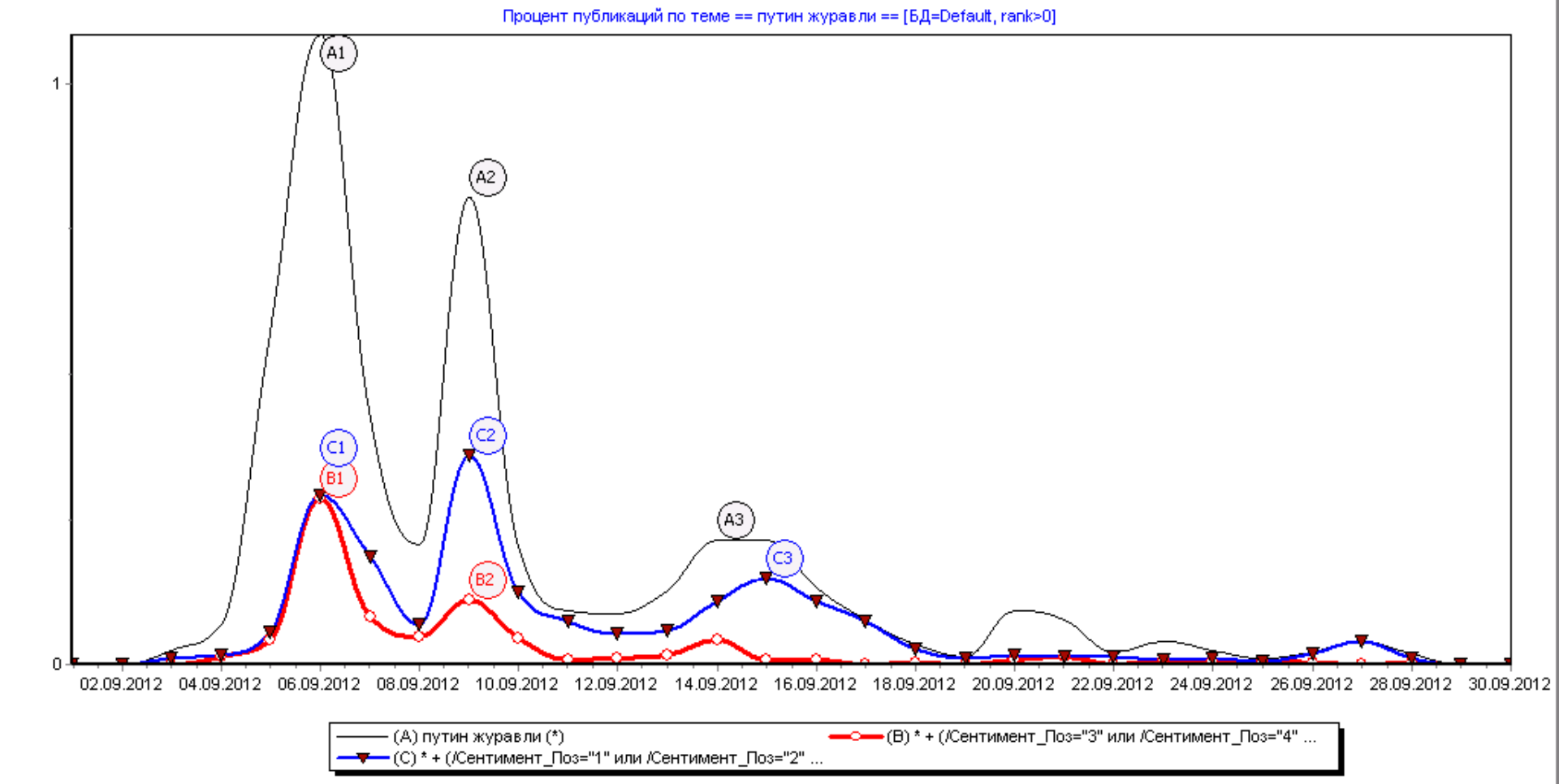
30.09.2014 08:15:02 **ВИЦЕ-МЭРА ЧЕБАРКУЛЯ ЗАПОДОЗРИЛИ В ПОЛУЧЕНИИ ВЗЯТКИ [DOSTUP1.RU]**

Вице-мэра Чебаркуля заподозрили в получении взятки 30 сентября 2014 года, 10:15Вице-мэра Чебаркуля заподозрили в получении взяткиЧЕЛЯБИНСК, АН "Доступ" В отношении заместителя главы Чебаркульского городского округа (Челябинская область) возбуждено уголовное дело – Алексея Сафонова подозревают в получении взятки, передает корреспондент Агентства новостей «Доступ». + 5 фрагм.

01.10.2014 10:04:58 **ГУБЕРНАТОР «ПОДКЛЮЧИЛ» СИЛОВИКОВ? [MOSCOW-POST.COM]**

Алексей Сафонов, вице-мэр Чебаркуля После того, как следователи не смогли выйти на связь с Сафоновым, чтобы задать ему дополнительные вопросы, появилось подозрение, что заместитель главы администрации Чебаркуля может попытаться сбежать. Однако, затем силовики все-таки смогли задержать Сафонова. Взятки на дорогах?

Sentiment Dynamics (*Putin and Siberian cranes*)



A1 06.09.2012 9:45:00 История о том, как президент РФ летал с птичьей стаей, в заголовках СМИ //1Рег_Алтапресс
A1 06.09.2012 12:41:00 Инопресса о Путине и журавлях «Трюки мачо, которому исполнилось 60» //slon.ru
A1 06.09.2012 13:35:00 Владимир Путин - вожак стаи журавлей. Как это было //0dp.ru
A2 09.09.2012 13:23:00 Путин гордится, что за ним полетели только самые сильные журавли //NewsMe.com.ua
A2 09.09.2012 13:58:00 Путин рассказал о слабых журавлях в своей стае //0ПОЛИТ.РУ
A2 09.09.2012 17:52:00 Собчак Мой Твиттер цитирует Путин на АТЭС //Аргументы Недели - статьи
A3 14.09.2012 8:02:00 Уволенная из-за полета Путина и журавлей журналистка Маша Гессен нашла новую работу //ИД "Собеседник"

(209,66)(307,147)(470,370)(209,343)(307,409)(209,323)(307,315)(502,395)

Sentiment. Aggregating user reviews

Positive

Приятное обслуживание [Pleasant service]

Отличное меню, вкусное пиво, приятное обслуживание
[Excellent menu, delicious beer, pleasant service]

Приличные порции [Nice portions]

Вкусная еда, приличные порции, хорошие напитки, доброжелательное обслуживание.
[Delicious food, decent portions, good drinks, benevolentservice]

Большой экран [Large screen]

Можно футбол смотреть набольшом экране.
*[Can watch football on a large screen]**…***

Negative

Хамством встретились впервые [With such rudeness met for the first time!!!]

С таким хамством встретились впервые!!!

Очень прокурено [Very smoky]

Очень прокурено.

Таракан [Cockroach]

Настоятельно \рекомендую\ посетить данное заведение, при условии, что вам нравится компания мышей и тараканов…
[We strongly \recommend\ to attend this place, provided that you like the company of mice and roaches…]

Забегаловка [Eatery]

Зашарпанная забегаловка.
[Dirty eatery]

Информация | Дерево | Кластеры | Лог построения рубрикатора

Отчет по запросу: мгу суперкомпьютеры /САНТИМЕНТ="+"

Формирование отчета: по кластерам
Рубрицирование по тезаурусу: НАУКА (Расширение: L) [K: 50]

АЭРОДИНАМИКА

(0.5) 10.08.2012 11:00:00 **Садовничий рассказал Путину про суперкомпьютер и космические спутники, которые запускает университет** **[Накануне.RU] #8520503#**
Садовничий рассказал Путину про суперкомпьютер и космические спутники, которые запускает университет Суперкомпьютер, работающий в МГУ, помог создать новые глазные капли и расшифровать американскую криптографию, сообщил ректор МГУ Виктор Садовничий на встрече с президентом России Владимиром Путиным, которая, как сообщили Накануне.RU в пресс-службе Кремля, состоялась в четверг. [51сп; Cluster: 5]

ЭЛЕКТРОНИКА

(0.67) 18.04.2012 14:14:00 **В МГУ появится новая магистерская программа по супервычислениям** **[0Российская газета - RG.RU] #5149389#**
В США, Германии, Франции, Китае вычислительные центры стали национальными. Сейчас наука на таком этапе развития, когда многие открытия невозможны без мощной вычислительной базы, - подчеркнул он на семинаре в МГУ, где ведущие ученые университета рассказывали о своих последних достижениях. Причем, все эти достижения были получены с помощью суперкомпьютера, которым владеет МГУ. [75сп; Cluster: 1]

(0.62) 26.09.2012 00:14:26 **МГУ имеет все возможности включиться в петафлопную гонку** **[0Независимая Газета] #10001590#**
Гордостью и украшением центра должен стать новый суперкомпьютер. Предполагается, что многие разработки будут вестись на средства инвесторов. Новый Ломоносовский корпус МГУ предназначен для ведущих научных коллективов, занимающихся перспективными исследованиями, и призван стать «технологическим поясом» университета. [75сп; Cluster: 2]

(0.52) 09.08.2012 21:14:27 **Ректор МГУ Виктор Садовничий сообщил Президенту о ключевых проектах в работе ВУЗа** **[0Первый Канал (ОРТ - видео)] #8506530#**
Ректор МГУ Виктор Садовничий сообщил Президенту о ключевых проектах в работе ВУЗа Важным направлением Виктор Садовничий назвал работу на суперкомпьютерах - этим заняты 600 научных коллективов. В частности, благодаря этим... [33сп; Cluster: 3]

(0.5) 10.08.2012 01:00:00 **Виктор Садовничий - Владимиру Путину: «Наш суперкомпьютер оказался в три раза мощнее, чем задумывали»** **[Комсомольская правда] #8510405#**
Виктор Садовничий - Владимиру Путину: «Наш суперкомпьютер оказался в три раза мощнее, чем задумывали» В четверг Владимир Путин провел встречу с ректором МГУ Виктором Садовничим, в ходе которой обсудил программу развития крупнейшего вуза страны, рассчитанную до 2020 года. [41сп; Cluster: 4]

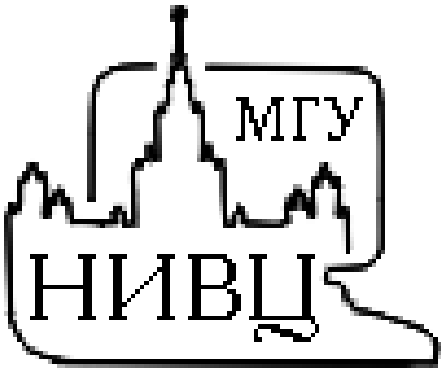
(0.5) 28.03.2012 07:35:00 **Top50 самых мощных компьютеров СНГ** **[0Ferra.ru - Компьютерные новости] #4517591#**
Научно-исследовательский вычислительный центр МГУ имени М.В.Ломоносова и Межведомственный Суперкомпьютерный Центр РАН выпустили шестнадцатую редакцию списка Top50 самых мощных компьютеров СНГ. Ломоносов Шестнадцатая редакция списка продемонстрировала

Выводы

- Задачи корпоративного информационного поиска достаточно многочисленны и разнообразны
- Высокопроизводительная обработка текстов возможна на уровне небольшой лаборатории/фирмы
- Основная проблема – плохо определенные задачи, трудности формирования обучающих множеств
- Достаточно большой процент ошибок даже для базовых технологий
- Одним из выходов является вовлечение человека в интерактивный диалог путем использования продвинутых средств визуализации
- Платой является резкий рост необходимых вычислений, необходимость задействования существенных вычислительных мощностей on-line

Conclusions

- The tasks of enterprise search are quite varied
- High-performance processing of corporate level texts corpora is possible on base of a small laboratory/company
- The main problem is badly defined objectives, difficulties of formation of the training sets
- A large percentage errors even for basic technologies
- Possible sollution is to involve the analyst in an interactive dialogue through the use of advanced visualization tools
- The fee is a burst increase in necessary computations, having to use substantial computational capacity on-line


МГУ
НИВЦ