Система мониторинга и анализа неструктурированной информации из разнородных источников

5 декабря 2012

Предпосылки

Высокие темпы развития Интернет ресурсов и обсуждение важных общественно-политических тем в блогосфере и социальных сетях создают условия для развития соответствующих инструментов анализа, которые позволят оперативно отслеживать изменяющийся информационный фон, что в свою очередь позволит повысить эффективность оперативных, так и стратегических управленческих планов Заказчика.

© 2011 SAP AG. All rights reserved.

Цель внедрения системы

Цель внедрения системы на первом этапе – получение полнофункционального, гибко настраиваемого инструмента для мониторинга и анализа неструктурированной информации из Интернет и других источников для оценки внешнего информационного фона Компании.

В дальнейшем планируется развитие системы для анализа внутренних и внешних данных на предмет выявления тональностей и скрытых связей.

© 2011 SAP AG. All rights reserved.

Пользователи системы

✓ **Руководство Компании – получение сводной аналитической информации, содействующей принятию стратегических решений.**

✓ **Специалисты по административной поддержке бизнеса – анализ взаимоотношений с государственными и регулирующими органами.**

✓ **Специалисты по корпоративным коммуникациям – анализ образа Компании в глазах общественности (через мониторинг СМИ и других источников), анализ узнаваемости брендов Компании и отношения к ним общественности.**

✓ **Специалисты по экономической безопасности – получение дополнительной информации о взаимосвязях и отношениях организаций-партнеров и отдельных персон, возможность отслеживания других взаимосвязей.**

✓ **Маркетологи – получение информации по целевым рынкам реализации продукции.**

✓ **Специалисты по стратегическому развитию Компании.**

© 2011 SAP AG. All rights reserved.

Источники информации

- **СМИ в сети интернет**
- **Социальные сети (Facebook, twitter, vkontakte…)**
- **Внутренние документы (договора, служебные записки)**
- **Корпоративная почта**
- **Система корпоративного документооборота**

© 2011 SAP AG. All rights reserved.

Основная концепция

Более 130+ Миллионов различных ресурсов

Экстракция сущностей

- Глубокий парсинг
- Грамматический анализ
- Индустриальная лексика

Хранилище данных

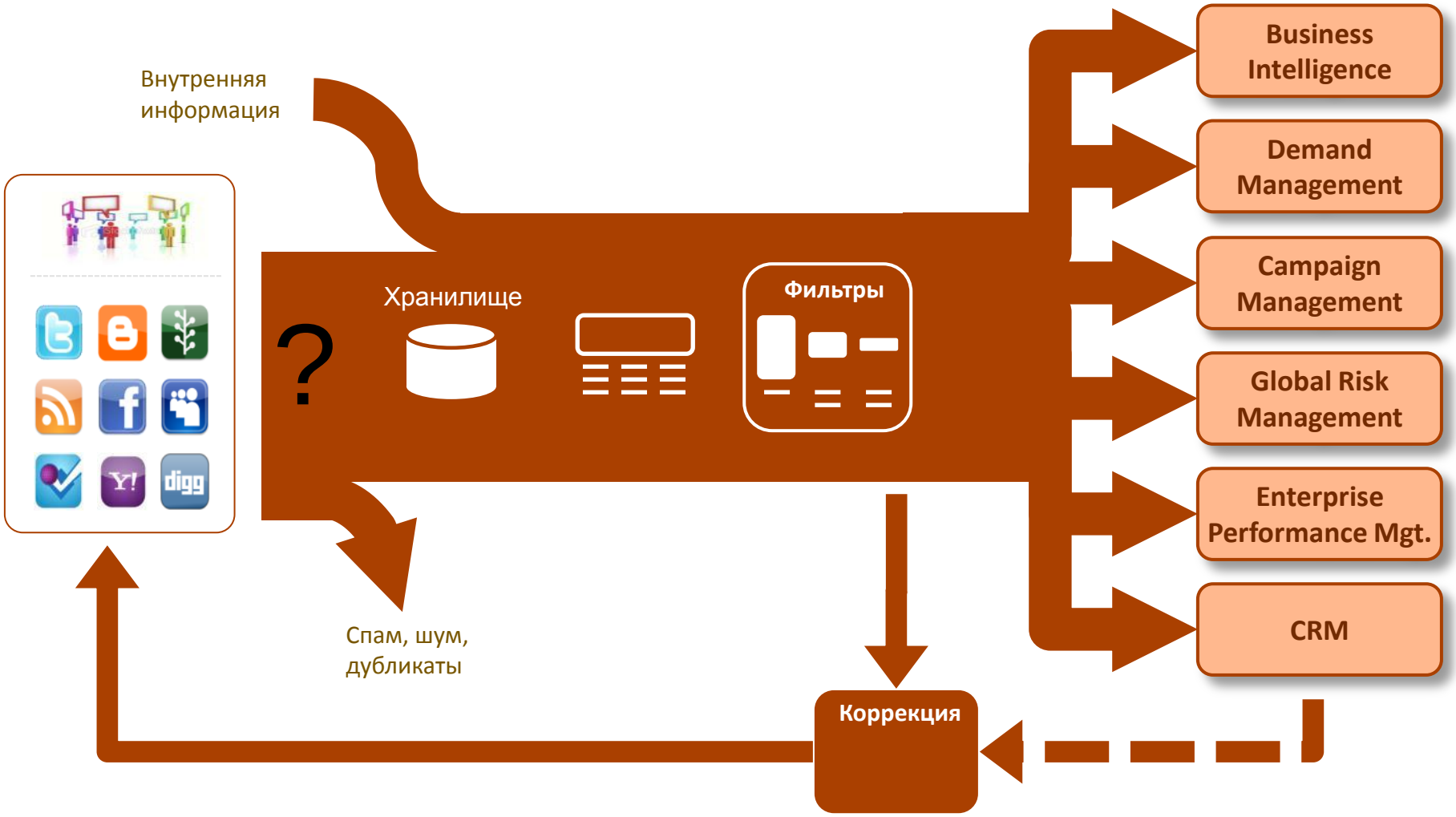
- SAP HANA
- SAP Sybase IQ

Готовые Панели управления и отчеты

Поиск и навигация

© 2011 SAP AG. All rights reserved.

Как это работает?



© 2011 SAP AG. All rights reserved.

Text Data Processing – работа с неструктурированной информацией

Основано на проверенной технологии

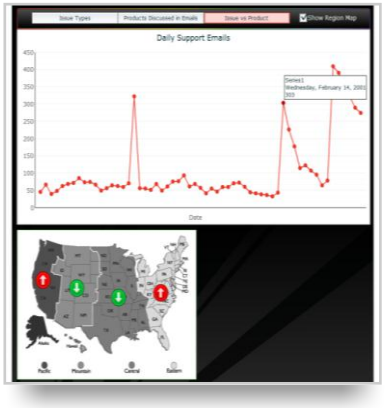
- Лидирующая на рынке технология Обработки Естественного Языка (NLP) для 31 языка – точная и производительная
- Основывается на **15+** годах исследований и технологиях Xerox PARC и Inxight Software до их приобретения BusinessObjects (2007)

© 2011 SAP AG. All rights reserved.

Text Data Processing (TDP)

1. Выделение сущностей
2. Очистка и сопоставление

Структурированная база данных



Информация доступна для…

Запросов

Анализа

Построения отчетности

Визуализации

Выявляет из источников текста ключевую информацию для дальнейшего анализа

© 2011 SAP AG. All rights reserved.

Архитектура Text Data Processing (TDP)

Разработчик ETL

Data Services Designer

Аналитик

Инструменты бизнес-аналитики

Семантический слой

Text Data Processing (извлечение сущностей)

Data Quality

Data Services

Источники

Сущности, факты

Целевое хранилище информации

© 2011 SAP AG. All rights reserved.

Что умеет Text Data Processing?

Text Data Processing извлекает из текста два типа элементов:

- **Базовые сущности:**

  Пол Маккартни был одним из Битлз.

  <PERSON>Пол Маккартни</PERSON> был одним из Битлз.

- **Домены:**

  Мне нравится ваше устройство.

  Мне <STRONGPOSITIVESENTIMENT>нравится</SPS><TOPIC> ваше устройство </TOPIC>.

© 2011 SAP AG. All rights reserved.

Извлечение базовых сущностей

Извлечение базовых сущностей часто называют *Named Entity Recognition (NER)* [*распознавание поименованных сущностей*].

Named entity recognition (NER) стремится обнаружить и классифицировать атомарные элементы текста, используя предопределенные категории, такие как имена людей, названия организаций, мест, выражений времени, количества, денежных сумм, процентов и т.п.

Системы NER получают на входе неаннотированный блок текста – например, такой:

Джим купил 300 акций Acme Corp. в 2006 году.

и формируют аннотированный блок текста, подобный следующему:

***<PERSON>**Джим**</PERSON>** купил **<QUANTITY>**300**</QUANTITY>** акций **<ORGANIZATION>**Acme Corp.**</ORGANIZATION>** в **<DATE>**2006 году**</DATE>**.*

-- Википедия

© 2011 SAP AG. All rights reserved.

Как работает извлечение базовых сущностей

Это не поиск по ключевым словам. Text Data Processing применяет лингвистические и статистические методы для выделения нужных сущностей.

- **Грамматический разбор**
  - *Can we **bill** you?*
  - ***Bill** was the president.*
- **Семантическая многозначность**
  - *I talked to **Bill** yesterday.*
  - *The duck has a **bill**.*
  - *The **bill** was signed into law.*

© 2011 SAP AG. All rights reserved.

Поддерживаемые типы сущностей

Сейчас	Планируется
Люди	Продукты (товары)
Должности	Даты
Организации коммерческие	Год, Месяц, День
Организации прочие	Временные интервалы
Страны	Праздники
Города (места)	Адрес
Номера телефонов	Регион
E-mail	Национальности
URL	Языки

© 2011 SAP AG. All rights reserved.

Извлечение доменов (фактов)

Извлечение доменов реализовано на основе извлечения базовых сущностей при помощи правил, выделяющих факты – отношения между сущностями или состояниями сущностей.

- **Отношение:**

  Василий прочитал книгу.

- **Состояние:**

  День рождения Марины 13 июля.

© 2011 SAP AG. All rights reserved.

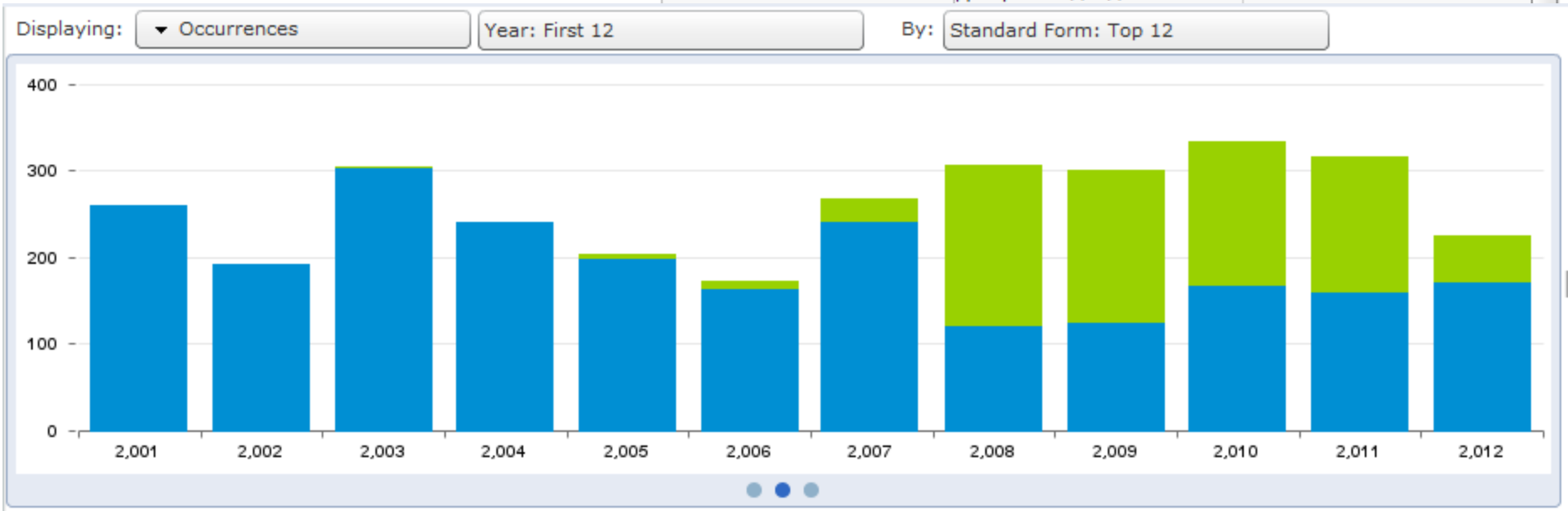
Извлечение доменов (фактов)

Примеры доменов
Смена руководства компаний (назначения, отставки)
Членство в организациях и партиях
Слияния и поглощения
Информация об организации (основатель, местоположение, контактная информация, …)
Вывод на рынок новых продуктов
События, связанные с терроризмом
Внешний облик человека
Взаимоотношения людей (в семье, коллективе, обществе)
Голос клиента (отношение к продукту/услуге)

© 2011 SAP AG. All rights reserved.

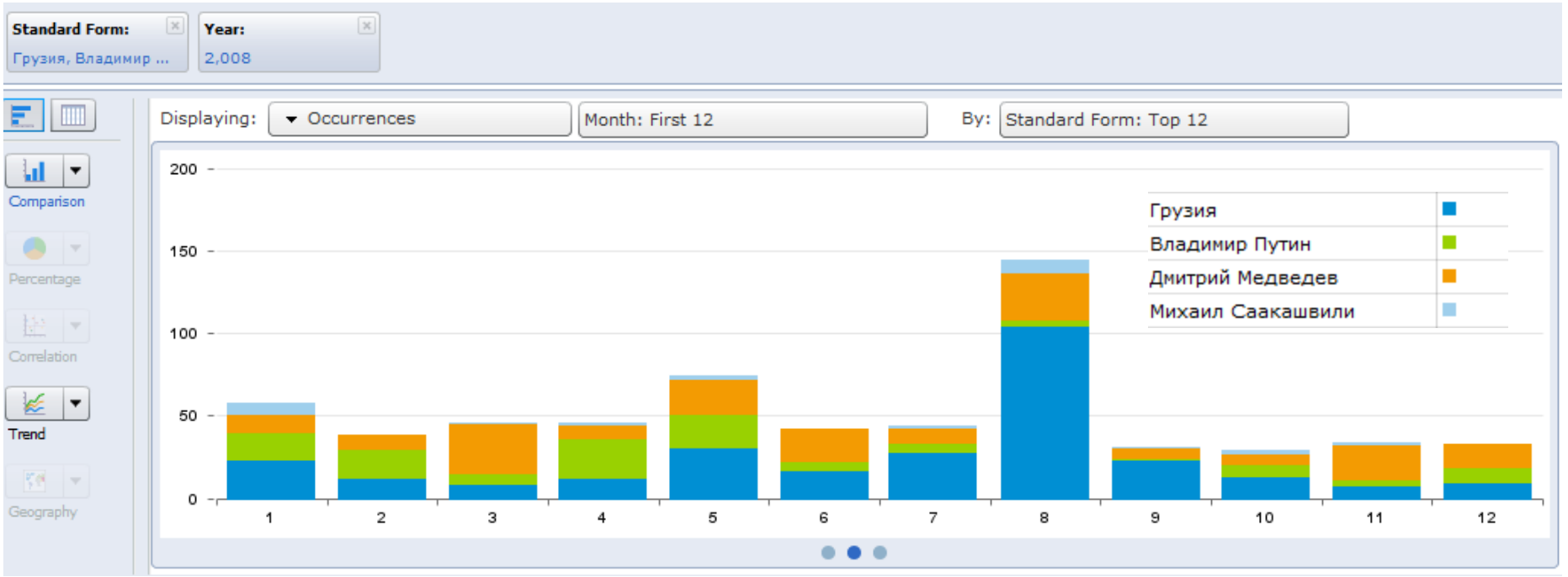
Пример анализа неструктурированных данных в SAP BusinessObejcts Explorer (выборы)

Year	Standard Form	Occurrences
2,001	Владимир Путин	260
	Дмитрий Медведев	
2,002	Владимир Путин	191
	Дмитрий Медведев	1
2,003	Владимир Путин	303
	Дмитрий Медведев	1
2,004	Владимир Путин	241
	Дмитрий Медведев	



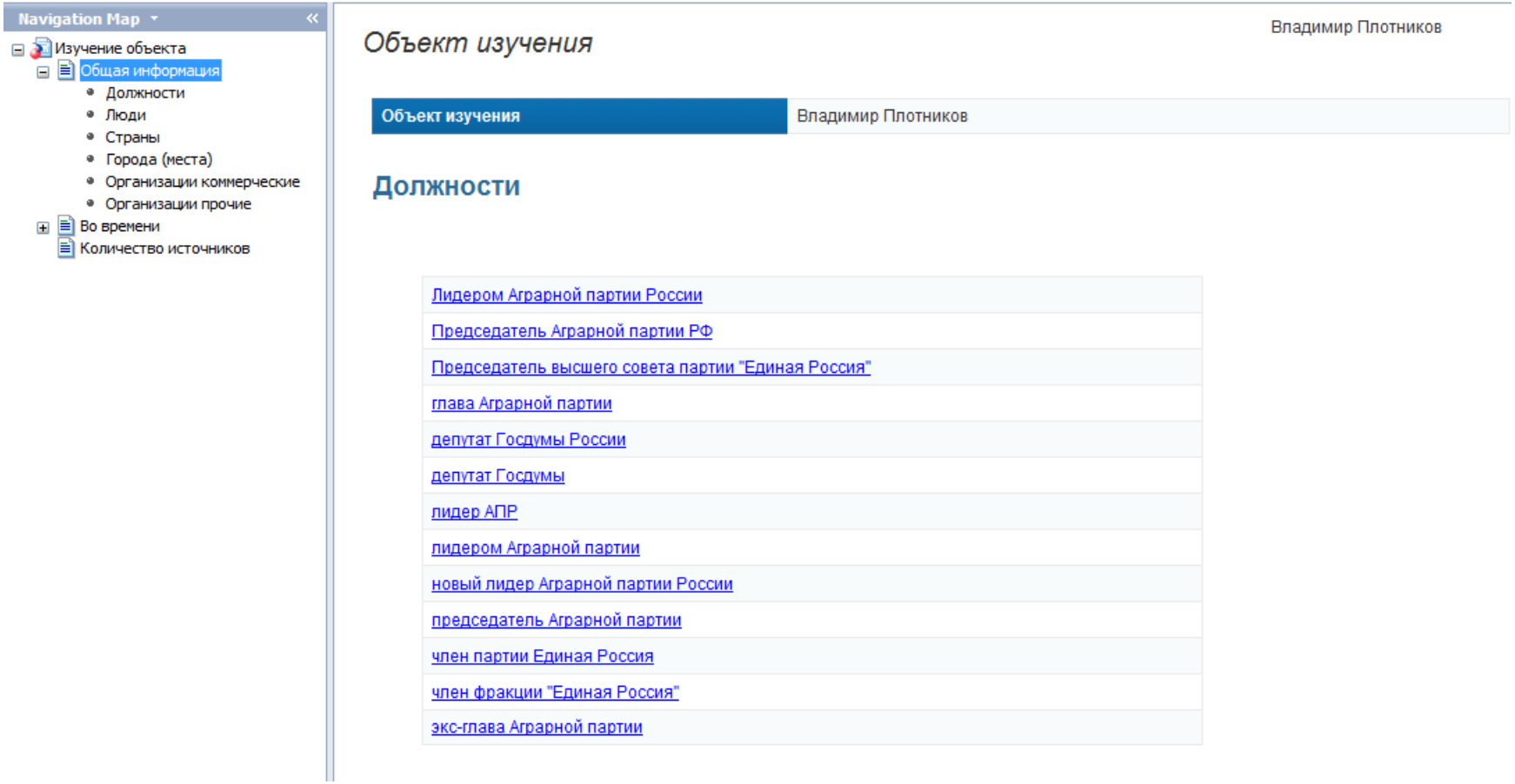
© 2011 SAP AG. All rights reserved.

Пример анализа неструктурированных данных в SAP BusinessObjects Explorer (война в Грузии)



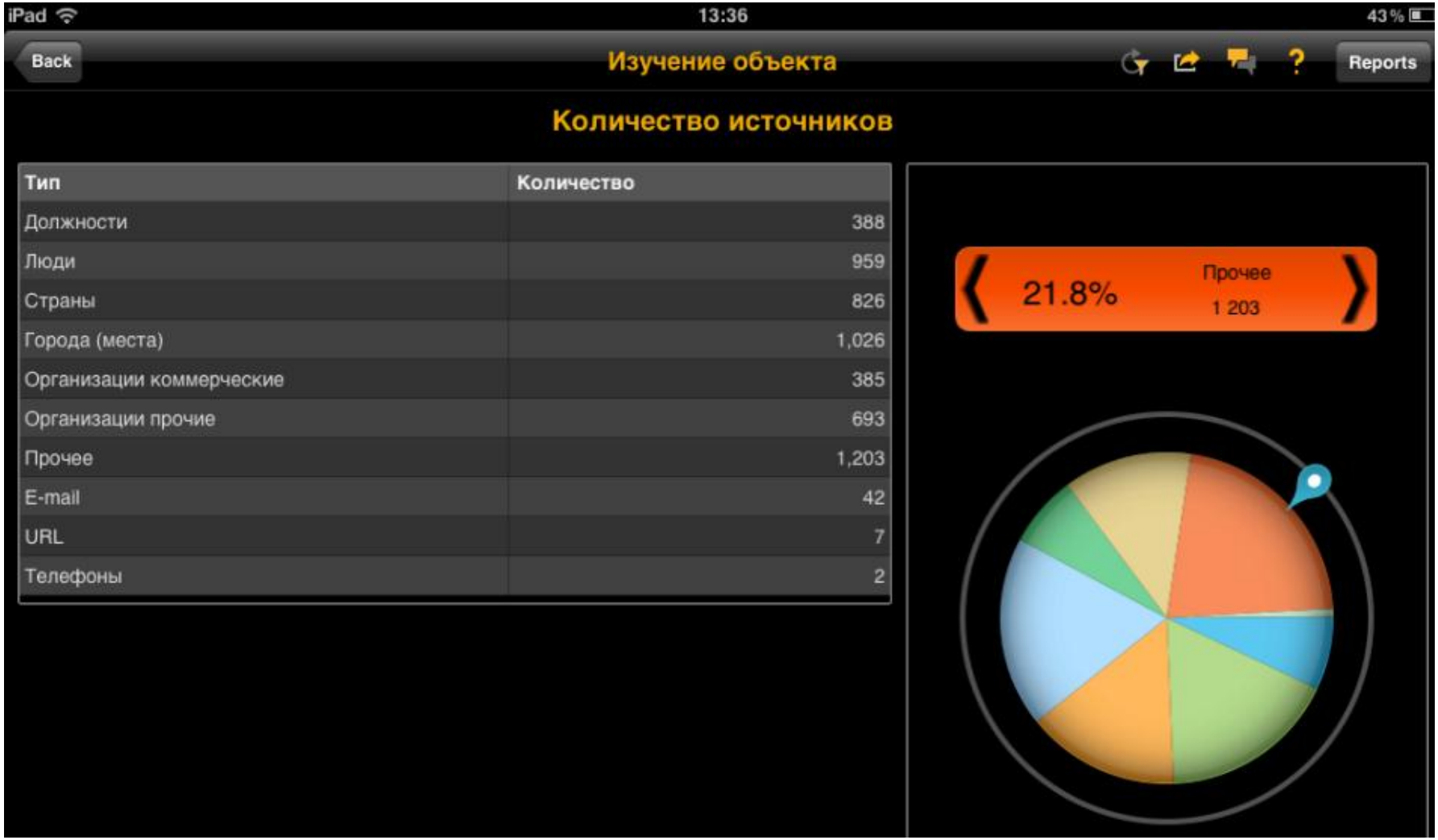
© 2011 SAP AG. All rights reserved.

Пример анализа неструктурированных данных в SAP BusinessObejcts WebIntelligence



© 2011 SAP AG. All rights reserved.

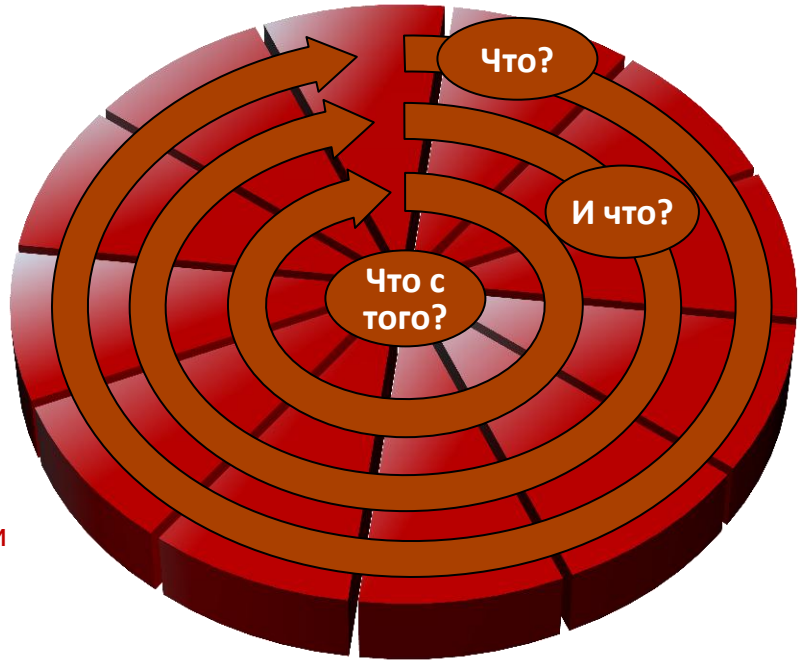
Пример анализа неструктурированных данных в SAP BusinessObejcts BI Mobile



© 2011 SAP AG. All rights reserved.

Для чего это можно использовать?

1. Анализ узнаваемости бернда
2. Отслеживание Маркетинговых компаний
3. Вывод на рынок новой продукци
4. Ранжирование/ отслеживание влияния
5. Социальный маркетинг
6. Влияние на выбор
7. Забота о клиенте
8. Управление рисками
9. Анализ деятельности конкурентов
10. Мониторинг партнеров
11. Категорийный анализ
12. Инновации



© 2011 SAP AG. All rights reserved.

Спасибо за внимание

© 2011 SAP AG. All rights reserved.