МОДЕЛИ, МЕТОДЫ И ПРОГРАММНЫЕ ИНСТРУМЕНТЫ ПОИСКА В СТРУКТУРНО РАЗМЕЧЕННЫХ ТЕКСТАХ

А.М. Гусенков
Казанский федеральный университет
2016

BIG DATA (БОЛЬШИЕ ДАННЫЕ)

Характеристики:

- **большой объем информации (Volume)**
- **разнообразие форматов данных разной степени структурированности : базы данных, текстовые документы, графические изображения, аудио-видео файлы (Variety) (главный критерий Big Data)**
- **изменчивость информации (Variability)**
- **скорость накопления и обработки данных (Velocity)**

Технологии: NoSQL, MapReduce, Hadoop, SAP HANA

ОБЪЕКТЫ ИССЛЕДОВАНИЯ BIG DATA

- **РБД, структурно размеченные своими схемами. Разметка РБД, связанная с нормализацией таблиц, предназначена для минимизации дублирования и поддержки целостности данных**

- **Полнотекстовые естественнонаучные документы, содержащие математические выражения (формулы). Математические выражения размечены для их графического отображения**

Цель: разработка методик и построения поисковых систем на естественном языке с использованием существующих разметок.

УРОВНИ ИНТЕГРАЦИИ

- **Физический** - конверсия данных из различных источников в единый формат их физического представления
- **Логический** - доступ к данным, содержащимся в различных источниках, в терминах единой глобальной схемы, которая описывает их совместное представление
- **Семантический** - обеспечивает поддержку единого представления данных с учетом их семантических свойств в контексте единой онтологии предметной области.

Достоинство семантического подхода:

- высокоуровневая модель данных является основой пользовательского интерфейса
- возможность рассуждений в терминах онтологии выступает в качестве концептуальной модели.

ОПРЕДЕЛЕНИЕ ОНТОЛОГИИ

Т.А. Гаврилова, В.Ф. Хорошевский:

Под формальной моделью онтологии O будем понимать упорядоченную тройку вида:

$$O = < X, R, Ф >,$$

где

- **X – конечное множество концептов (понятий, терминов) предметной области, которую представляет онтология O;**
- **R – конечное множество отношений между концептами (понятиями, терминами) заданной предметной области;**
- **Ф – конечное множество функций интерпретации (аксиоматизация), заданных на концептах и/или отношениях онтологии O.**

СТРУКТУРНЫЕ ПРОБЛЕМЫ РБД

- **Различия в физической структуре таблиц баз данных, т.е. в распределении столбцов по таблицам.**
- **Различия в уровнях абстракции при проектировании баз данных.**
  - **На уровне таблиц: структуры данных типа, имеющего несколько субтипов, могут быть представлены одной таблицей или отдельными таблицами для каждого из субтипов.**
  - **На уровне столбцов: вынос значения столбца, входящего в состав ключа, в наименование столбца.**
  - **Комбинированный случай: в наименование столбцов или таблиц выносится комбинация ключевых параметров.**
- **Неатомарность атрибутов: атрибут, имеющий сложную структуру, может быть представлен как одним столбцом, так и различными наборами столбцов.**
- **Шкалирование атрибутов: однотипные атрибуты представлены в разных единицах измерения.**
- **Отсутствие атрибутов: значение подразумевается по умолчанию.**
- **Различия в наименовании объектов БД: синонимы, сокращения, аббревиатуры, различающиеся грамматические конструкции.**

ЛЕКСИКО-СЕМАНТИЧЕСКИЕ ПРОБЛЕМЫ

Типичное описание таблицы

Идент. столбца	Описание столбца
NC	Номер скважины
GOD	Год
MES	Месяц
PL	Код пласта
SPEX	Способ эксплуатации
DN	Добыча нефти
DW	Добыча воды
DG	Добыча газа
KDEX	Часов работы
PLB	Плотность попутно добытой воды

Состав исследуемых баз данных

	MS SQL	Oracle1	Oracle2
Таблиц	644	219	95
Столбцов	11688	2028	524

Общее количество терминов в исследуемых базах данных

Записи	Кол-во
Общее количество	48 629
Из них из Oracle	24 035
Из них из MS SQL	24 594

СОПОСТАВЛЕНИЕ ТАБЛИЦ БД

Столбцы таблицы 1	Лексико-семантическое отношение	Столбцы таблицы 2
Номер скважины	**Меронимия** (составная часть)	**Объект разработки**
Код пласта		
Год	**Гипонимия** (частный случай общего понятия)	**Период эксплуатации**
Месяц		
Способ эксплуатации	**Синонимия**	**Метод разработки**
Добыча воды	**Конверсия** (обратное отношение к субъекту действия), **Гипонимия**	**Тип флюида**
Добыча нефти		**Отдача флюида**
Добыча газа		
Часов работы	**Антонимия**	**Процент простоя скважины**
Плотность попутно добытой воды		

ПРЕДСТАВЛЕНИЕ РБД В ФОРМАЛИЗМЕ ОНТОЛОГИЙ

Онтология O

Универсальные концепты (C):

ТАБЛИЦА, СТОЛБЕЦ, КЛЮЧ, ДОМЕН,

Универсальные отношения (R):

- **ТАБЛИЦА содержит СТОЛБЕЦ**
- **ТАБЛИЦА имеет первичный КЛЮЧ**
- **ТАБЛИЦА имеет внешний КЛЮЧ**
- **КЛЮЧ содержит СТОЛБЕЦ**
- **СТОЛБЕЦ имеет тип ДОМЕН**

ИЗВЛЕЧЕНИЕ АТРИБУТОВ

Правила РБД разрешают соединения таблиц только по общему ключу, причем в одной из таблиц этот ключ играет роль первичного, а в другой – вторичного ключа.

- **ФИ1 (соединение по ключу):**

Если ТАБЛИЦА1 имеет первичный КЛЮЧ1 и ТАБЛИЦА2 имеет внешний КЛЮЧ1, то существует ТАБЛИЦА3, содержащая столбцы, принадлежащие ТАБЛИЦА1 и ТАБЛИЦА2.

- **ФИ2 (проекция реляционного отношения для сокращения множества столбцов до искомого):**

Если ТАБЛИЦА1 содержит СТОЛБЕЦ1, то существует ТАБЛИЦА2, содержащая все остальные столбцы ТАБЛИЦА1, кроме СТОЛБЕЦ1.

ПОСТРОЕНИЕ ОНТОЛОГИИ ПРЕДМЕТНОЙ ОБЛАСТИ

Модель данных

Epicentre версии 3.0 нефтетехнической корпорации Petrotechnical Open Software Corporation (POSC). Представляется в виде ER-диаграмм, а также в виде набора текстовых файлов на языке EXPRESS (ISO 10303, part 11). Более 1000 сущностей.

Средство представления онтологий

язык OWL от Semantic Web Activity, рекомендованный консорциумом W3C.

ОСНОВНЫЕ ПОНЯТИЯ МОДЕЛИ ДАННЫХ Epicentre

- **Сущность (entity)**
- **Атрибуты сущности (attributes): явные и инверсные**
- **Супертип и подтип (sypertype, subtype)**
- **Типы данных**
- **Ограничения согласованности и уникальности**

ОСНОВНЫЕ СТРУКТУРЫ ЯЗЫКА ОНТОЛОГИИ OWL

- **Простые именованные классы** (Class, subClassOf)
- **Индивиды**
- **Простые свойства** (ObjectProperty, DatatypeProperty)
- **Свойства индивидов**
- **Характеристики свойств** (TransitiveProperty, SymmetricProperty, FunctionalProperty, inverseOf, InverseFunctionalProperty)
- **Ограничения свойств** (allValuesFrom, someValuesFrom, кардинальность, hasValue)

КОНВЕРТАЦИЯ МОДЕЛИ Epicentre В OWL-DL

- **Модель данных Epicentre на EXPRESS – 500 страниц**
- **LR(1)-грамматика со встроенной семантикой – 30 страниц**
- **Реализация – Java, JavaCup**
- **Представление Epicentre на OWL – 3500 страниц**

ЛИНГВИСТИЧЕСКИЙ ТЕЗАУРУС

Словарь предметной области - словоформы из описания:

- **сущностей и атрибутов модели Epicentre**
- **атрибутов таблиц и доменов таблиц-справочников РБД реальной нефтедобывающей корпорации.**

Синсеты - входные синонимические ряды. Особенности синсетов: короткие фразы, сокращения, аббревиатуры.

Отношения на словоформах и синсетах: гипонимия, часть – целое, несовместимость, антонимия, конверсивность, омонимия.

ИНТЕГРАЦИЯ РБД НА ОСНОВЕ ОНТОЛОГИЙ

POSC

Типа WordNet – язык ПО

Универсальная для любых РБД

ПРЕДЛАГАЕМЫЙ ПОДХОД

- **Ведение диалога между системой и пользователем в табличном виде.**
- **Использование семантических подходов для поиска столбцов**
- **Автоматический поиск возможных соединений по ключам БД**
- **Использование визуальных процедур для задания операций селекции.**

ПРИНЦИП ОРГАНИЗАЦИИ ДИАЛОГА ТАБЛИЧНОЕ ЗАДАНИЕ ЗАПРОСА

Табличная форма

Номер скважины	10*
Дата ввода в эксплуатацию	>01.07.1995
Дата КРС	<01.01.2001
Дебит нефти ожидаемый	>0
Дебит нефти фактический	
Стоимость ремонта	>10 млн. руб.

Текстовая форма

?

ЧЕЛОВЕКО-МАШИННЫЙ ДИАЛОГ

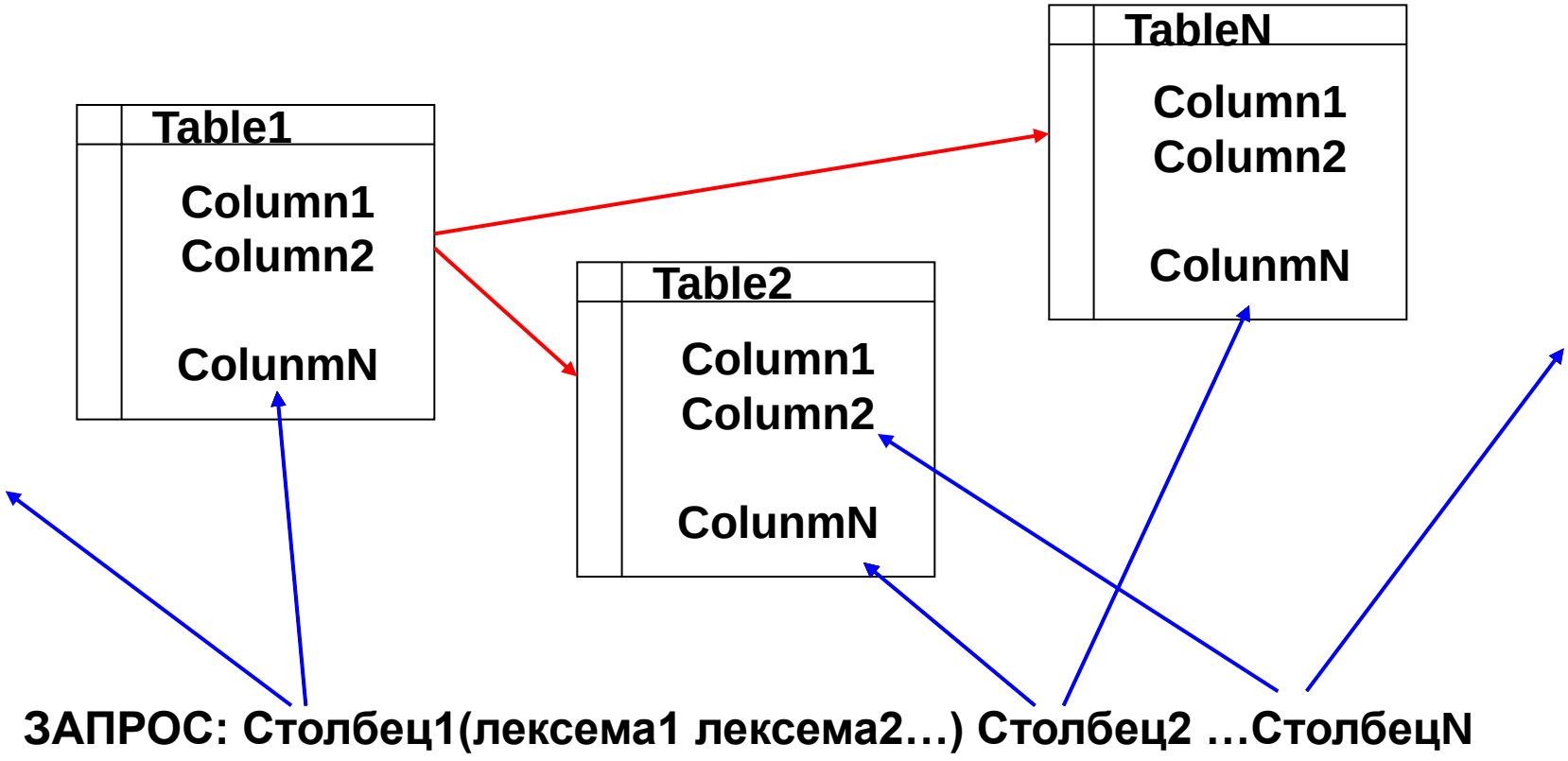
Пользователь

- Конструирует таблицу
- Уточняет формулировки
- Выбирает нужные столбцы при наличии альтернатив
- Задает дополнительные параметры запроса, запускает на выполнение
- Получает ответ в виде таблицы

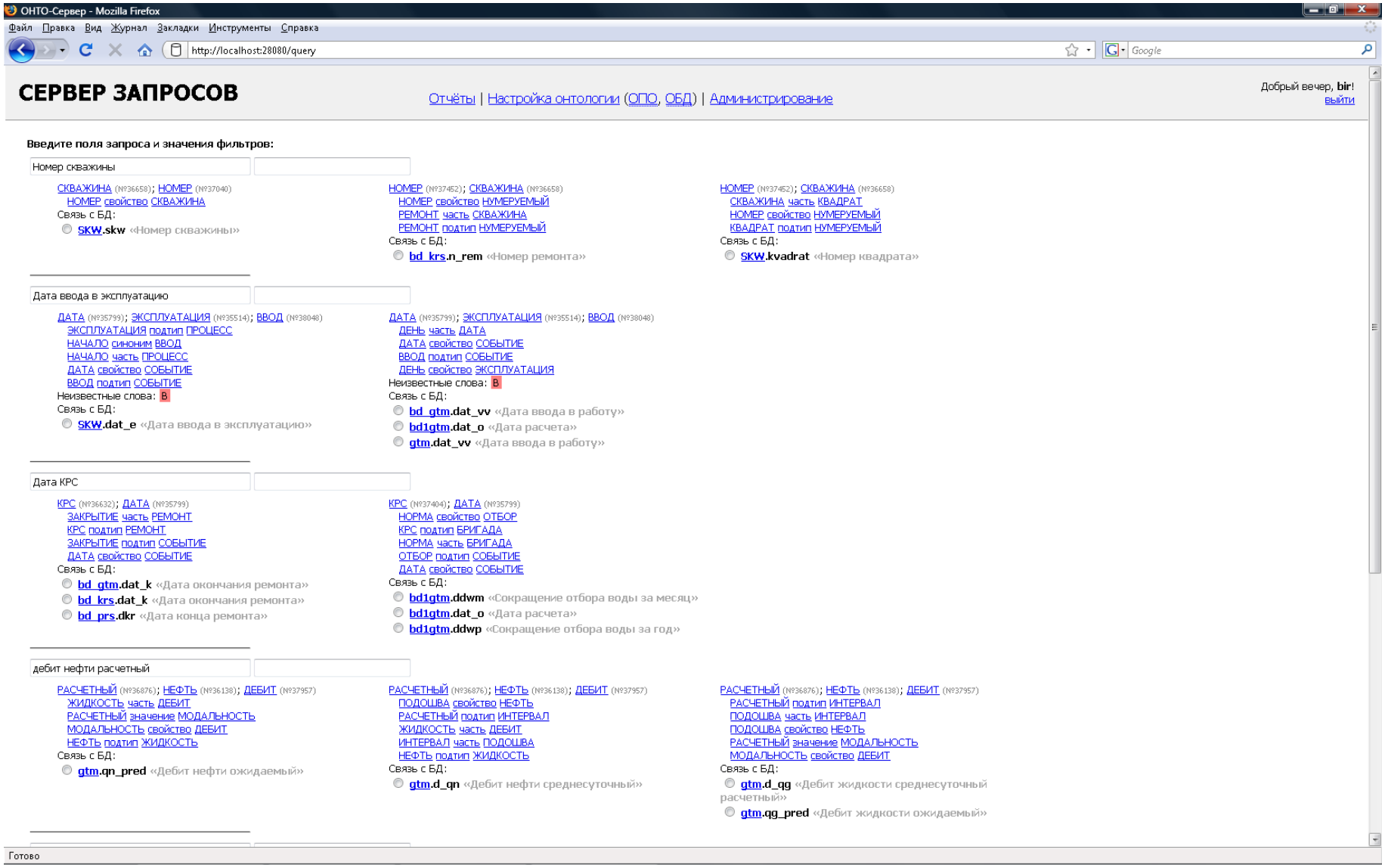
Система

- «Понимает» запрос
- Сопоставляет элементы - столбцы запроса со столбцами БД
- Генерирует SQL-запрос
- Выполняет запрос
- Запоминает контекст пользователя

ГЕНЕРАЦИЯ SQL
ПОИСК СОЕДИНЕНИЙ В БД



ОБЪЯСНЕНИЕ ОТВЕТА СИСТЕМЫ



ИНСТРУМЕНТЫ НАСТРОЙКИ СИСТЕМЫ

- **Редактор языка предметной области (словаря)**
- **Редактор онтологии предметной области**
- **Редактор онтологии базы данных**
- **Настройки алгоритмов поиска**
- **Анализ полноты и адекватности онтологий.**

ПОДХОДЫ К РАЗМЕТКЕ

Поисковый объект:

сложный нелинейный нетекстовый объект (формула и её переменные)

Подходы:

1. **Индексация документов по ключевым словам и связывание с ними формульных выражений. Использован для поиска формул в Википедии.**
2. **Включение формульных выражений в онтологию и связывание их с терминами онтологии для организации поискового запроса на естественном языке.**

КОНТЕКСТ ФОРМУЛЫ

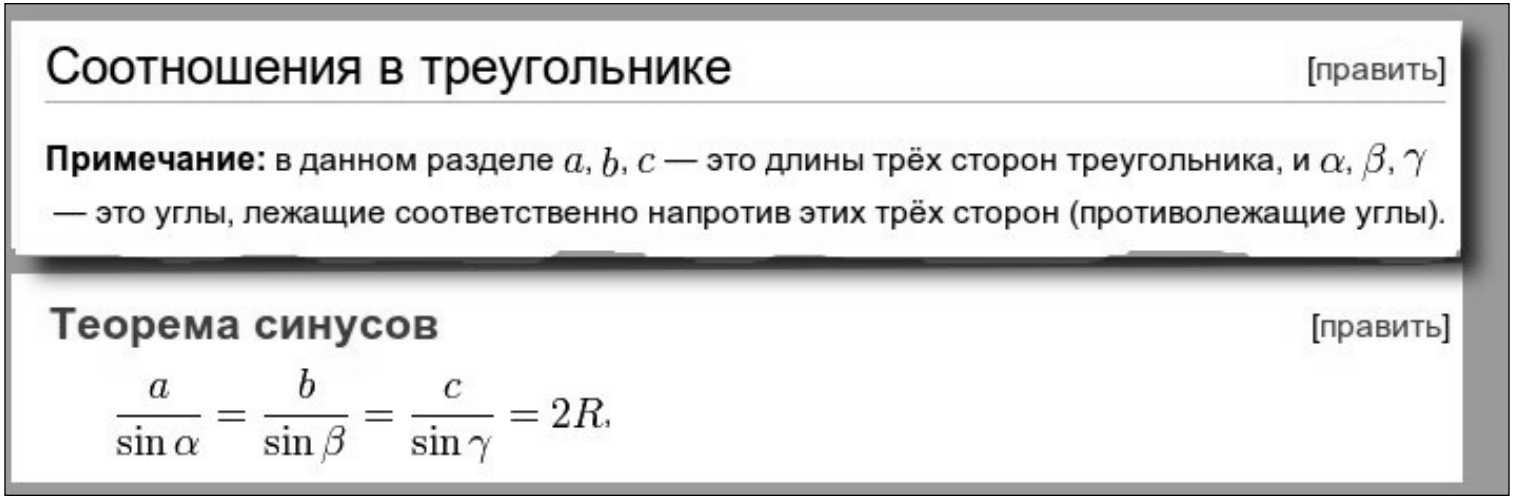
Сущности:

- естественнонаучные термины,
- символьные условные обозначения терминов (переменные)
- математические фрагменты (формулы).

Отношения:

- «термины - переменные» (текстовое определение значения символа с помощью терминов)
- «переменные – формулы» (вхождение символа в формулу)

ПРИМЕР ФОРМУЛЬНОГО КОНТЕКСТА



Соотношения в треугольнике [править]

Примечание: в данном разделе a, b, c — это длины трёх сторон треугольника, и α, β, γ — это углы, лежащие соответственно напротив этих трёх сторон (противолежащие углы).

Теорема синусов [править]

$$\frac{a}{\sin\alpha} = \frac{b}{\sin\beta} = \frac{c}{\sin\gamma} = 2R,$$

Фрагмент статьи Википедии о площади треугольника

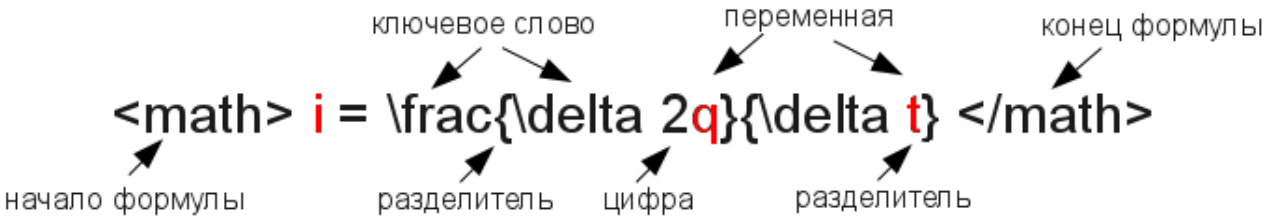
Определения переменных даны в сплошном тексте.

Формула - нетекстовый объект.

Предполагается, что появление текстового наименования переменной в окрестностях её символьного представления указывает на семантическую связь между ними.

МЕТОД РАЗМЕТКИ МАТЕМАТИЧЕСКИХ ФОРМУЛ

- **выделение формульных фрагментов**
- **классификация: формула или переменная**
- **связывание формул с переменными**



Результат:

Отношение много-ко-многим между формулами и переменными.

ПОИСК ФОРМУЛ В ВИКИПЕДИИ

Подсистемы системы поиска:

- **Загрузка и анализ данных Википедии;**
- **Полнотекстовое индексирование (Apache Lucene);**
- **Позициональное индексирование математических формул и переменных;**
- **Взаимодействие с пользователем (Web-приложение с доступом через браузер);**
- **Поиск и ранжирование**

ПОИСК И РАНЖИРОВАНИЕ

Первый этап:

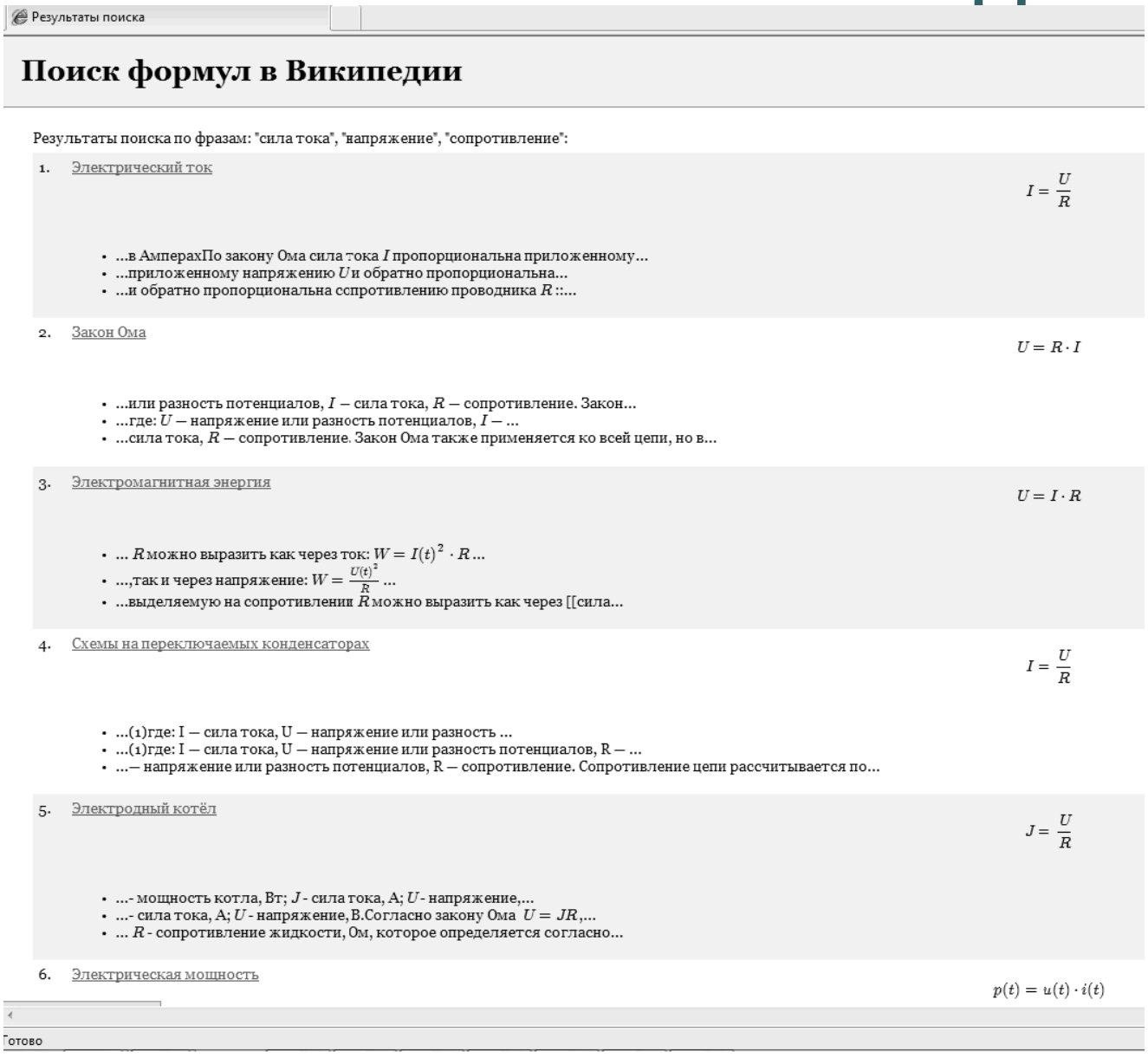
- **Полнотекстовый поиск всех вхождений терминов в тексты.**
- **Поиск переменной в некоторой окрестности термина с учётом МДР.**
- **Определение формулы, соответствующей переменной.**
- **Построение для формулы группы текстовых фрагментов, включающих термины и переменные.**

Второй этап:

- **Поиск наилучшей группы текстовых фрагментов.**

Все возможные сочетания полученных текстовых фрагментов проверяются по критерию близости (минимум СКО определений переменных от позиции формулы)

ПОИСК ФОРМУЛ В ВИКИПЕДИИ



Результаты поиска

Поиск формул в Википедии

Результаты поиска по фразам: "сила тока", "напряжение", "сопротивление":

1. Электрический ток

$$I = \frac{U}{R}$$

- ...в АмперахПо закону Ома сила тока I пропорциональна приложенному...
- ...приложенному напряжению U и обратно пропорциональна...
- ...и обратно пропорциональна сспротивлению проводника R ::...

2. Закон Ома

$$U = R \cdot I$$

- ...или разность потенциалов, I — сила тока, R — сопротивление. Закон...
- ...где: U — напряжение или разность потенциалов, I — ...
- ...сила тока, R — сопротивление. Закон Ома также применяется ко всей цепи, но в...

3. Электромагнитная энергия

$$U = I \cdot R$$

- ... R можно выразить как через ток: $W = I(t)^2 \cdot R$...
- ...,так и через напряжение: $W = \frac{U(t)^2}{R}$ ...
- ...выделяемую на сопротивлении R можно выразить как через [[сила...

4. Схемы на переключаемых конденсаторах

$$I = \frac{U}{R}$$

- ...(1)где: I — сила тока, U — напряжение или разность ...
- ...(1)где: I — сила тока, U — напряжение или разность потенциалов, R — ...
- ...— напряжение или разность потенциалов, R — сопротивление. Сопротивление цепи рассчитывается по...

5. Электродный котёл

$$J = \frac{U}{R}$$

- ...- мощность котла, Вт; J - сила тока, А; U - напряжение,...
- ...- сила тока, А; U - напряжение, В.Согласно закону Ома $U = JR$,...
- ... R - сопротивление жидкости, Ом, которое определяется согласно...

6. Электрическая мощность

$$p(t) = u(t) \cdot i(t)$$

Готово

Результаты поиска

закона Ома

по трем параметрам:

сила тока

напряжение

сопротивление

РАЗМЕТКА ДЛЯ ПОИСКА ПО ОНТОЛОГИИ

Цель разметки: связывание формул с именными группами

Обработка документа:

- выделение и анализ формульных фрагментов;
- определение связей между переменными и формульными фрагментами;
- определение связей между формульными фрагментами и именными группами;
- дополнение аннотаций Math атрибутами формульной разметки.

Реализация:

плагин к текстовому процессору GATE на языке Java; для разметки используются средства работы с аннотациями библиотеки Gate и оригинальные алгоритмы.

РАЗМЕТКА XML-ДОКУМЕНТА

- **Разметка стандартными аннотациями: Token, Sentence, Math и др. (Gate)**
- **Разметка аннотациями именных групп: TERM**
- **Построение на основе аннотаций Math внутренней модели документа, содержащей набор разобранных, классифицированных, связанных между собой формульных фрагментов.**
- **Дополнение аннотаций Math и TERM атрибутами связывания.**

СТАТИСТИКА СВЯЗЫВАНИЯ В ЗАВИСИМОСТИ ОТ МДР

Проведена ручная оценка качества связывания на двух корпусах математических текстов.

Анализ результатов связывания проведен на 8 документах при изменении МДР от 15 до 40 симметрично в обе стороны от формулы.

Всего:
1247 формул
1357 именных групп

МДР	Math	Terms	VirOK%	NotVirOK%	TotalOk%	VirBad%	Others%	TotalBad%
15	1247	1357	36,33	30,47	66,80	23,90	9,30	33,20
20	1247	1357	42,34	25,50	67,84	25,66	6,50	32,16
25	1247	1357	40,98	20,69	61,67	23,02	15,32	38,33
30	1247	1357	41,38	21,49	62,87	27,83	9,30	37,13
35	1247	1357	41,86	21,01	62,87	29,03	8,10	37,13
40	1247	1357	42,02	19,65	61,67	29,67	8,66	38,33

VirOK – правильное связывание

NotVirOK – правильное несвязывание (математическое выражение находится в контексте, не содержащем его определения)

VirBad – связывание математического выражения в контексте, не предполагающем связывания

Научная новизна

- Построено представление структуры РБД в формализме онтологий
- Предложена методика построения онтологии предметной области на основе логической модели «сущность–связь», представленной в виде ER-диаграмм
- Разработана методика нерегламентированного доступа к РБД
- Предложен метод разметки математических выражений для организации семантического поиска в естественнонаучных текстах, содержащих математические выражения

Результаты, выносимые на защиту

- Модель интеграции РБД в формализме онтологий для предметной области, связанной с нефтедобычей
- Методика построения онтологии предметной области на основе логической модели «сущность–связь», представленной в виде ER-диаграмм
- Методика и программная реализация системы для выполнения нерегламентированного доступа к РБД на естественном языке в терминах предметной области
- Метод разметки математических выражений для организации семантического поиска в естественнонаучных текстах, содержащих математические выражения, и программная реализация системы семантического поиска математических выражений в статьях Википедии и системы разметки естественнонаучных текстов, содержащих математические выражения, для поиска по онтологии

Спасибо за внимание!