На правах рукописи

Гусенков Александр Михайлович

МОДЕЛИ, МЕТОДЫ И ПРОГРАММНЫЕ ИНСТРУМЕНТЫ ПОИСКА В СТРУКТУРНО РАЗМЕЧЕННЫХ ТЕКСТАХ

Специальность 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата технических наук

г. Казань –2016

Работа выполнена на кафедре технологий программирования Института вычислительной математики и информационных технологий Казанского (Приволжского) федерального университета.

Научный руководитель: доктор физико-математических наук, профессор Казанского (Приволжского) федерального университета, Заслуженный деятель науки Республики Татарстан
Елизаров Александр Михайлович

Официальные оппоненты: доктор технических наук, профессор Уфимского государственного авиационного технического университета, Заслуженный деятель науки Российской Федерации
Куликов Геннадий Григорьевич

кандидат технических наук, доцент, заведующий лабораторией научных электронных информационных ресурсов Института проблем рынка Российской Академии наук
Когаловский Михаил Рувимович

Ведущая организация: Казанский национальный исследовательский технический университет (КНИТУ-КАИ) им. А.Н. Туполева

Защита состоится 23 декабря 2016 года в 14.30 часов на заседании диссертационного совета Д 212.081.35 в Казанском (Приволжском) федеральном университете по адресу: 420008, г. Казань, ул. Кремлевская, д. 35, ауд. 1011.

С диссертацией можно ознакомиться в Научной библиотеке им. Н.И. Лобачевского Казанского (Приволжского) федерального университета.

Автореферат разослан _____ 2016 года

Ученый секретарь диссертационного совета
кандидат физ.-мат. наук, доцент Еникеев А.И.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Экспоненциальный рост накопленных массивов данных, объём которых в настоящее время измеряется в зеттабайтах, привел к качественным изменениям в IT-технологиях сбора, хранения, управления, обработки и анализа информации. Одновременно в научный оборот вошло понятие Big Data (Большие Данные). Например, международная исследовательская компания Forrester[1] определяет это понятие как технологию в области аппаратного и программного обеспечения, которая объединяет, организует, управляет и анализирует данные, характеризующиеся «четырьмя V»: объемом (Volume), разнообразием (Variety), изменчивостью (Variability) и скоростью (Velocity):

- Volume – очень большой объем информации, накопленный в базах данных, трудоемко обрабатывать и хранить традиционными средствами СУБД;
- Variety – разнообразие форматов данных (главный критерий Big Data): большие массивы данных, поступающие из разных источников в различных форматах, разной степени структурированности – табличные данные в СУБД, иерархические данные, текстовые документы, видео, изображения, аудиофайлы и т. д.;
- Variability – изменчивость информации: например, информация, непрерывно поступающая с датчиков некоторых устройств или из интернета, имеющая важное значение для анализа, прогнозирования и принятия решений;
- Velocity – скорость накопления и обработки данных; в ряде задач востребованы технологии обработки данных в реальном времени.

В настоящее время разработаны технологии работы с Big Data, наиболее известными из них являются следующие:

- NoSQL – ряд подходов к реализации хранилищ баз данных, отличных от РБД, для работы с данными, структура которых не может быть жестко определена;
- MapReduce – модель распределения вычислений, используется для параллельных вычислений при обработке очень больших наборов данных;
- Hadoop – фреймворк с набором утилит и библиотек для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов; система защищена от выхода из строя любого из узлов кластера путем дублирования данных на других узлах;
- SAP HANA – высокопроизводительная NewSQL-платформа для хранения и обработки данных; сочетание технологий OLAP и OTLP обеспечивает как высокоскоростную обработку транзакций, так и работу со сложными аналитическими запросами в рамках единой платформы.

[1] http://www.asterdata.com/newsletter-images/30-04-2012/resources/forrester_expand_your_digital_horiz.pdf

Проблема интеграции гетерогенных электронных ресурсов чрезвычайно многоаспектна и многообразна. М.Р. Когаловский [2] предложил следующую классификацию систем интеграции данных: интеграция информации на физическом, логическом и семантическом уровнях. Интеграция на физическом уровне сводится к конверсии данных из различных источников в требуемый единый формат их физического представления. Интеграция на логическом уровне предусматривает возможность доступа к данным, содержащимся в различных источниках, в терминах единой глобальной схемы, которая описывает их совместное представление с учетом структурных и, возможно, поведенческих свойств данных. Интеграция на семантическом уровне обеспечивает поддержку единого представления данных с учетом их семантических свойств в контексте единой онтологии предметной области. Достоинство семантического подхода заключается в том, что основой пользовательского интерфейса является при этом высокоуровневая модель данных, а возможность рассуждений в терминах онтологии выступает в качестве концептуальной модели. В качестве средства описания онтологий в диссертации используются формализм теории графов и язык OWL[3], разработанный рабочей группой Semantic Web Activity и рекомендованный консорциумом W3C.

Одной из основных целей интеграции ресурсов является возможность организации эффективного поиска информации в интегрированных электронных ресурсах.

В настоящее время активно развиваются исследования в области компьютерной лингвистики и обработки информации, представленной на естественном языке. Это работы, связанные с созданием электронных словарей, тезаурусов, онтологий, а также алгоритмы автоматического извлечения фактов из текста. В рамках этого направления разработано большое количество специализированных систем поиска в различных предметных областях.

В диссертации представлен подход к интеллектуальному поиску сложных объектов в массивах больших данных (Big Data). Исследованы два типа представления информационных объектов: реляционные базы данных (РБД), которые структурно размечены своими схемами, и полнотекстовые естественнонаучные документы, содержащие математические выражения (формулы). Для таких полнотекстовых документов предложена дополнительная автоматизированная разметка для организации поиска формул. В обоих случаях источником информации для построения онтологии и дальнейшей организации поиска являются тексты на естественном языке, которые относятся к слабоструктурированным данным. Для РБД это комментарии к наименованиям таблиц и их атрибутов, а для естественнонаучных документов (статей, монографий и т. д.) – текстовое содержимое размеченных документов. Для эффективной интеграции реляционных баз данных использованы информация, извлекаемая из самой

[2] http://www.ipr-ras.ru//articles/kogalov10-05.pdf
[3] https://www.w3.org/TR/2004/REC-owl-features-20040210/

базы данных, а также более общая информация, относящаяся к предметной области в целом. Таким образом, для успешного решения задачи интеграции РБД использованы вспомогательные информационные ресурсы, содержащие физические модели баз данных, логические модели предметной области и тезаурусы пользовательской терминологии, представленные в формализме онтологий.

Предложенный подход оказался применимым к задаче конструирования нерегламентированных запросов к наборам баз данных в больших информационных системах, насчитывающих несколько десятков локальных баз данных с различными логической структурой и физической организацией, но относящихся к одной предметной области. Этот подход апробирован и в другой предметной области – при поиске математических выражений (формул) в естественнонаучных текстах по наименованиям (определениям) переменных, входящих в формулу, так как формула фактически является отношением между своими переменными.

Примеры реализации поиска сложных объектов в структурно размеченных массивах данных, рассмотренные в диссертации, показали возможность применения комбинаторного подхода поиска путей на графах с использованием известных приемов минимизации перебора. При этом не использовалось никакой априорной семантической информации о предметной области. Более того, данный подход применялся к задачам, совершенно разным как по семантике, так и по формам представления данных, что позволяет говорить о достаточной универсальности предложенного подхода.

Цель и основные задачи. Цель диссертации состоит в разработке методик интеграции и построения поисковых систем на естественном языке для дальнейшего применения в гетерогенных структурно размеченных слабоструктурированных массивах больших данных (Big Data), в том числе, с использованием существующих разметок. Разметка РБД, связанная с нормализацией таблиц, в первую очередь, предназначена для минимизации дублирования и поддержки целостности данных. В естественнонаучных текстах математические выражения (формулы) размечены для их графического отображения.

Основные задачи:

- анализ структурных проблем интеграции РБД; разработка структуры представления РБД в формализме онтологий;
- создание онтологии и построение лингвистического тезауруса языка предметной области;
- разработка и программная реализация нерегламентированного доступа к РБД на естественном языке в терминах предметной области;
- разработка методов разметки естественнонаучных текстов, содержащих математические выражения; реализация алгоритмов семантического поиска математических выражений в статьях Википедии, размеченных разработанным методом;

- разработка и реализация методов разметки естественнонаучных текстов для поиска по онтологии.

Объектом диссертационного исследования являются реляционные базы данных, слабоструктурированные текстовые описания, представление знаний в формализме онтологий и поисковые системы.

Предмет исследования. Методы семантической интеграции разнородных структурированных текстовых описаний на основе онтологий.

Методы исследования основаны на теории реляционных баз данных, теории формальных языков и грамматик, алгоритмов теории графов. Кроме того, использованы методы, разработанные в области интеграции данных, информационного поиска, машинного обучения и онтологического инжиниринга, описанные в работах отечественных и зарубежных ученых: Д.А. Поспелова, Т.А. Гавриловой, А.С. Нариньяни, В.Ф. Хорошевского, Г.С. Осипова, М.Р. Когаловского, С.Д. Кузнецова, Н. Гуарино, Н. Ной, Т. Грубера, Т. Бернерса-Ли, Д. МакГиннесс, Ф. Баадера, Д. Фенселя и др.

Научная новизна диссертационной работы состоит в следующем:

- 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.

Практическая значимость результатов диссертации. Концепция интеграции РБД на основе взаимосвязанных онтологий, программная реализация нерегламентированного доступа к РБД и набор инструментальных средств для работы с онтологиями могут использоваться и получить дальнейшее развитие в нефтедобывающих компаниях. Кроме того, развитый подход может быть применен и для других предметных областей, представленных ER-диаграммами.

Предложенный метод разметки математических выражений в естественнонаучных текстах и реализованный на его основе алгоритм поиска математических формул показал достаточно хорошую релевантность результатов поиска в сочетании с его высокой скоростью. В диссертационной работе этот метод был использован при разметке математических выражений для поиска по онтологии. Предложенный метод может найти применение при создании других поисковых систем, связанных с нахождением формул.

Достоверность полученных результатов обеспечена применением строгого математического аппарата теории реляционных баз данных, методов теории формальных языков и грамматик, алгоритмов теории графов, а также экспериментальной проверкой работоспособности и эффективности прикладных

программных систем, реализованных на основе предложенных методик и методов.

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.

Апробация результатов работы. Результаты диссертации докладывались на следующих конференциях:

- Республиканский научный семинар АН РТ «Методы моделирования» (Казань, 18 ноября 2015 г.);
- Казанские школы по компьютерной и когнитивной лингвистике (Казань) TEL-2014 (6–9 февраля 2014 г.), TEL-2012 (25-28 января 2012 г.), TEL-2008 (10–13 декабря 2008 г.), TEL-2006 (9–11 декабря 2006 г.); TEL-2005 (8–10 декабря 2005 г.);
- Третья международная конференция по когнитивной науке (Москва, 20–25 июня 2008 г.);
- Всероссийская конференция с международным участием «Знания–Онтологии–Теории» (ЗОНТ-07) (Новосибирск, Институт математики им. С.Л. Соболева СО РАН, 14–16 сентября 2007 г.);
- Международная конференция Диалог-2007 (Бекасово, 30 мая – 3 июня 2007 г.);
- 9-ая Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL-2007» (Переславль-Залесский, 15–18 октября 2007 г.);
- 3-я ежегодная международная научно-практическая конференция «Инфокоммуникационные технологии глобального информационного общества» (Казань, 8–9 сентября 2005 г.),

а также на ежегодных Итоговых научных конференциях и ряде кафедральных семинаров Казанского (Приволжского) федерального университета (2006–2016 гг.).

Диссертационное исследование проводилось при поддержке грантов РФФИ 06-07-89219 «Разработка методики интеграции реляционных баз данных

на основе онтологий» и 11-07-00507 «Методы и технологии извлечения, представления, интеллектуального анализа многоуровневой логической структуры связной коллекции научных публикаций».

Результаты диссертационного исследования внедрены в нефтегазодобывающей компании ПАО «Татнефть» и в образовательный процесс Казанского (Приволжского) федерального университета.

Содержание диссертации соответствует паспорту научной специальности 05.13.11 Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей, пункты 3 и 4.

Структура диссертации. Диссертация состоит из введения, трех глав, заключения, библиографии и трех приложений, содержит 139 страниц (111 страниц основного текста), 39 рисунков, 20 таблиц. Список литературы содержит 111 источников.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении описаны проблемы, рассмотренные в диссертации, обоснована актуальность исследования, сформулированы цели и задачи работы.

В первой главе дан обзор известных подходов к решению задач интеграции РБД. Исследованы особенности представления информации в РБД. Предложен способ интеграции гетерогенных РБД на основе онтологий и описано построение онтологий РБД, предметной области и лингвистического тезауруса.

В диссертации выделены следующие структурные проблемы интеграции РБД:

- различия в физической структуре таблиц баз данных (т. е. в распределении столбцов по таблицам), обусловленные гибкостью правил построения схем БД (так называемых правил нормализации); например, некоторая группа неключевых атрибутов может быть как сохранена в основной таблице, так и вынесена в отдельную таблицу;
- различия в уровнях абстракции при проектировании баз данных:
  - на уровне таблиц: структуры данного типа, имеющего несколько незначительно отличающихся субтипов, можно представить как отдельными таблицами для каждого из субтипов (сегментация по ключу), так и единой таблицей;
  - 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, когда атрибут, имеющий сложную структуру, может быть представлен как одним столбцом, так и различными наборами столбцов;
- шкалирование атрибутов, когда однотипные атрибуты представлены в разных единицах измерения;

- отсутствие атрибутов, значение которых подразумевается по умолчанию;
- различия в наименовании объектов баз данных: в наименованиях одних и тех же объектов могут быть использованы синонимы, сокращения, аббревиатуры, применены различающиеся грамматические конструкции.

Определение онтологии. В диссертации принято следующее определение онтологии (Т.А. Гаврилова, В.Ф. Хорошевский [4]): формальной моделью онтологии О является упорядоченная тройка 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**.

Онтология РБД. В диссертации введено понятие онтологии РБД (О), не зависящее от конкретного экземпляра РБД. Определены универсальные концепты (X) ТАБЛИЦА, СТОЛБЕЦ, КЛЮЧ, ДОМЕН, соответствующие основным объектам баз данных, и универсальные отношения между ними (R): ТАБЛИЦА содержит СТОЛБЕЦ; ТАБЛИЦА имеет первичный КЛЮЧ; ТАБЛИЦА имеет внешний КЛЮЧ; КЛЮЧ содержит СТОЛБЕЦ; СТОЛБЕЦ имеет тип ДОМЕН. Объекты (таблицы, столбцы, ключи и домены) конкретной базы данных в этом случае представляются как экземпляры универсальных концептов соответствующего типа.

Введены две функции интерпретации (**ФИ**):

ФИ1: Если ТАБЛИЦА1 имеет первичный КЛЮЧ1 и ТАБЛИЦА2 имеет внешний КЛЮЧ1, то существует ТАБЛИЦА3, содержащая столбцы, принадлежащие ТАБЛИЦА1 и ТАБЛИЦА2.

ФИ2: Если ТАБЛИЦА1 содержит СТОЛБЕЦ1, то существует ТАБЛИЦА2, содержащая все остальные столбцы ТАБЛИЦА1, кроме СТОЛБЕЦ1.

Первая функция интерпретации соответствует операции соединения по ключу, вторая – операции проекции реляционного отношения, необходимой для сокращения множества столбцов, получаемого при соединении таблиц, до искомого.

Таким образом, задача извлечения информации из интегрируемых РБД сводится к нахождению способов извлечения заданных атрибутов (столбцов) из таблиц РБД, т.е. к нахождению такой последовательности применения ФИ1 и ФИ2, которая даст в результате искомое множество столбцов {С} из онтологии О. Эта задача принадлежит известному классу задач о блуждании по ориентированному графу, где вершинами являются экземпляры концепта ТАБЛИЦА, а дугами – наличие общего ключа, ориентированного посредством отношений «имеет первичный» и «имеет вторичный». Подходы к решению данного класса

[4] Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2001. 384 с.

задач хорошо известны и представляют лишь вычислительную сложность при большой размерности задачи.

Онтология предметной области. В качестве прототипа онтологии предметной области выбрана логическая модель данных Epicentre нефтетехнической корпорации Petrotechnical Open Software Corporation (POSC[5]), имеющая статус отраслевого стандарта. Модель представлена в виде ER-диаграмм, а также набора текстовых файлов на языке EXPRESS. В модели данных Epicentre определено более 1000 объектов (entities), связанных с разведкой и добычей нефти. В модели определены характеристики, которые могут содержать сущности, названные атрибутами сущностей (attributes). Архитектура Epicentre основана на объектно-ориентированной концепции наследования классов.

В качестве средства представления онтологии выбран язык OWL-DL (Description Logic). Выбор языка обусловлен его поддержкой в существующих системах описания знаний и системах логического программирования. Разработана схема конвертации модели Epicentre в язык описания онтологий OWL. Применены следующие основные подходы:

- любой сущности Epicentre соответствует простой именованный класс OWL-онтологии с сохранением в именах классов приставок, позволяющих идентифицировать сущности-свойства и сущности-справочники; все эти классы располагаются в корне таксономического дерева онтологии;
- степени связности сущностей (один-к-одному, один-ко-многим, многие-ко-многим) в OWL соответствует определение простых свойств-атрибутов, если связанная сущность не является типом данных, и свойств-значений в противном случае; указание степени связи между классами реализовано с помощью понятия кардинальности в OWL.

В OWL отсутствуют структурные элементы, которые в полном объеме описывают определение уникальности Epicentre. Поэтому в определение каждого класса на языке OWL добавлено новое предопределенное свойство, в котором перечислены все атрибуты, образующие уникальный ключ. Аналогичным образом решена проблема сохранения условий ограничений.

Для каждой категории данных Epicentre в OWL-онтологии построены отдельные классы, в свойствах которых использовались встроенные типы данных языка OWL. Построена формальная LR(1)-грамматика модели Epicentre, на основе которой реализовано семантическое преобразование модели Epicentre, описанной на языке EXPRESS, в онтологию на языке OWL. Выполнена русификация описания сущностей и атрибутов модели Epicentre, а также соответствующих им классов и свойств на OWL.

Лексико-семантические проблемы. Идентификаторы столбцов реальных баз данных неинформативны для задачи идентификации их семантики. Более информативными являются определения столбцов, в которых обычно исполь-

[5] http://www.energistics.org/energistics-standards-directory/epicentre-archive

зуются лексика естественного языка и профессиональные термины из предметной области, что переводит задачу интеграции в область лингвистики.

Исследование лексико-семантических проблем интеграции проведено на реальных базах данных Публичного акционерного общества (ПАО) «Татнефть» при решении задачи переноса данных из базы данных MS SQL в две базы данных Oracle, в совокупности эквивалентных по составу информации в базе данных MS SQL (рисунок 1).

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

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

Общее количество терминов в исследуемых базах данных

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

Рисунок 1

Структуры исследуемых баз данных существенно различаются по количеству столбцов и таблиц. Между тем количества семантически различных терминов в этих базах данных близки (24035 и 24504).

Рисунок 2 демонстрирует многочисленные лексико-семантические отношения между определениями столбцов конкретных таблиц базы данных MS SQL и базы данных Oracle. Например, «добыча» и «отдача» являются конверсивами, выражающими точку зрения на процесс: «Скважина добывает нефть из залежи», «Залежь отдает нефть скважине». Соответствие столбцов «Часов работы» и «Процент простоя скважины» основано на антонимии терминов «Работа» и «Простой», к тому же выраженных в различных единицах (часы и проценты). Достаточно часто встречаются отношения метонимии типа «Объект – Роль», «Процесс – Результат» и некоторые другие.

Лингвистический тезаурус. Для создания лингвистического тезауруса природно-технических объектов использованы принципы построения тезаурусов WordNet. Словарь предметной области построен путем объединения словоформ из описаний сущностей и атрибутов модели Epicentre и описаний атри-

бутов таблиц и доменов таблиц-справочников РБД ПАО «Татнефть». Для каждого слова определен входной синонимический ряд (синсет). На лексико-семантических вариантах слов и синсетах определены следующие отношения: гипонимия, часть – целое, несовместимость, антонимия, конверсивность, омонимия.

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		

Рисунок 2. Пример сопоставления таблиц различных баз данных

Во второй главе приведен обзор существующих подходов к реализации нерегламентированного доступа к РБД и предложена методика автоматической генерации SQL-запросов, требующая от пользователя знаний только в своей предметной области. Рассмотрена реализация системы на основе предложенной методики.

Автоматизация построения онтологии предметной области и языка предметной области (лингвистического тезауруса), а также способ создания экземпляра онтологии произвольной РБД обуславливают применимость данного подхода в технологиях сбора, хранения и обработки информации больших данных.

Подход, предложенный в диссертации, базируется на следующих принципах:

- ведение диалога между системой и пользователем в табличном виде;
- Использование семантических подходов для поиска столбцов;
- использование визуальных процедур для задания операций селекции.

Пользователь формулирует запрос в виде таблицы на естественном языке, используя термины тезауруса. В таблице 1 приведен пример модельного запроса. Очевидно, что подобный запрос, сформулированный в виде предло-

жения на естественном языке, был бы весьма громоздким не только для машинного анализа, но и для понимания человеком.

Название столбца	Условие
Номер скважины	10*
Дата ввода в эксплуатацию	> 1.06.2001
Дата КРС	
Дебит нефти ожидаемый	>0
Дебит нефти фактический	

Таблица 1. Модельный запрос

Семантическая информация представлена в системе в виде семантической сети (онтологии) предметной области, которая содержит набор возможных связей между концептами. Операция семантического разбора комментариев к названиям столбцов РБД почти полностью автоматическая, что позволяет достаточно быстро привязать к предметной области большую корпоративную базу данных. Семантический разбор запроса пользователя заключается в выявлении подграфов на семантической сети, которые связывают концепты, именуемые словами и словосочетаниями, употребленными в названии столбцов. На искомые подграфы наложено условие минимальности либо по количеству связей, либо по их суммарной длине.

Поиск столбцов РБД, наиболее релевантных названиям, задаваемым пользователем в своем запросе, производится путем сопоставления подграфов семантического разбора столбцов запроса и столбцов РБД.

Для выявленных релевантных кандидатов, которые могут содержаться в различных таблицах, автоматически выявляются подмножества, для которых возможно построение соединений по ключам. Если возникает несколько комбинаций связанных релевантных столбцов, то дополнительно учитывается количество записей в построенном соединении.

Построенный запрос предъявляется пользователю для задания дополнительных условий в режиме визуального конструктора.

Реализация системы. Система реализована в виде веб-приложения с доступом через браузер. Знания SQL и структуры БД рядовому пользователю не требуются. Возможности по просмотру и редактированию SQL-запроса могут пригодиться программистам или опытным пользователям, которые хотят сконструировать сложный отчет из многих запросов. В этом случае SQL-операторы, сгенерированные системой, могут использоваться как шаблоны для доработки запросов в более профессиональных системах, например, Crystal Reports.

По запросу пользователя генерируются описание «понимания» системой введенной в каждом столбце фразы и ее возможные привязки к столбцам базы данных. Пользователь может проконтролировать правильность понимания запроса системой по предоставляемой ему информации:

- варианты выполненного системой разбора как подграф онтологии предметной области; пользователю представляется набор связей, найденных системой с учетом структуры онтологии и общей длины подграфа разбора;
- для каждого варианта разбора выводятся возможные варианты местонахождения их в РБД (таблица, столбец и комментарий к столбцу).

Сгенерированный запрос может быть отредактирован или сохранен и использован в качестве прототипа для построения новых запросов. Контекст запросов пользователя (выбранные таблицы, столбцы и связи между ними) сохраняется между сеансами работы и используется в последующем для ограничения перебора соединений таблиц и сокращения времени построения запроса.

Технологическая подсистема содержит редактор языка предметной области (словаря), редактор онтологии предметной области, редактор онтологии базы данных, настройки алгоритмов поиска, инструменты для анализа полноты и адекватности онтологий.

Разработанная система протестирована на РБД ПАО «Татнефть», содержавших порядка 700 таблиц и более 11 тысяч столбцов, тезаурус системы составлял около 6 тысяч слов, онтология предметной области содержала около 1.5 тысяч концептов. По большинству несложных запросов система показала достаточно хорошую релевантность поиска ответов на соответствующие запросы.

Скорость выполнения запроса сильно зависит от количества столбцов, задаваемых в нем. Основное время тратится на поиск связей столбцов в РБД. Таким образом, предлагаемая технология в текущей реализации эффективно работает с запросами, содержащими небольшое количество столбцов, что вполне достаточно для запросов справочного характера. Для ограничения вариантов перебора соединений таблиц используются связи, сохраненные в контексте пользователя, что позволяет повысить реактивность системы.

В третьей главе проанализирован подход к семантическому поиску математических формул по текстовым наименованиям входящих в них переменных. Приведен обзор существующих подходов. Предложен метод разметки естественнонаучных текстов для установления связей между текстовыми определениями переменных, их обозначениями и формулами, в состав которых эти переменные входят. Описаны две реализации систем поиска: поиск формул в статьях интернет-энциклопедии Википедия и разметка корпусов математических текстов для поиска по онтологии. Обсуждены результаты оценочных экспериментов с точки зрения релевантности поиска и полноты связывания, а также способы решения основных проблем предложенного подхода.

В качестве поискового объекта рассмотрен сложный нелинейный нетекстовый объект, включающий собственно математическое выражение формулы в нотации $L^{A}T_{E}X$ и набор определений символьных обозначений, участвующих в математическом выражении, которые извлекаются из всего анализируемого

текста. При этом в формулировке запроса использованы не математические конструкции, а словесные наименования переменных, входящих в искомую формулу.

В естественнонаучных текстах выделены следующие виды сущностей: **естественнонаучные термины**, **символьные условные обозначения терминов (переменные)**, **математические фрагменты (формулы)**.

Определим следующие отношения: **«термины – переменные»** и **«переменные – формулы».** Первое отношение есть текстовое определение значения символа в некотором контексте с помощью терминов, второе отношение указывает на вхождение символа в формулу. Предполагается, что появление текстового определения переменной в окрестностях её символьного представления указывает на семантическую связь между ними.

Все перечисленные сущности и отношения между ними составляют контекст формулы.

Пример расширенного формульного контекста приведен на рисунке 3, где определения переменных даны в сплошном тексте, а формула является нетекстовым объектом.

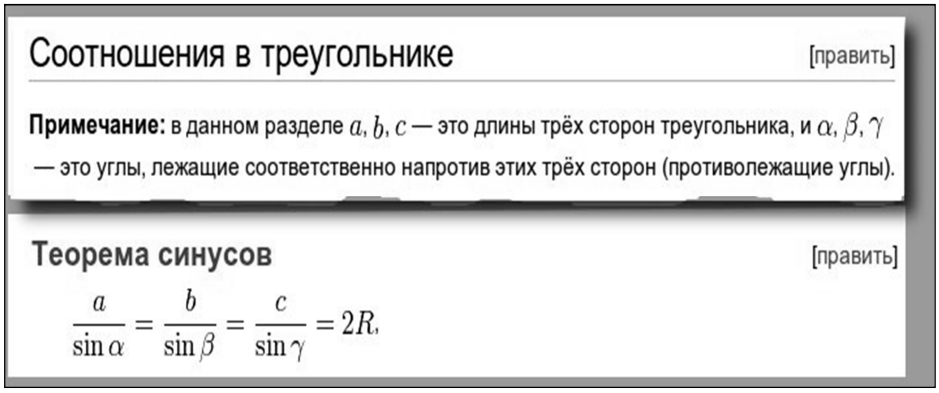

Рисунок 3. Пример формульного контекста (фрагмент статьи Википедии)

Метод разметки математических выражений

Шаг 1. Классификация математических выражений (МВ). МВ считается любой текст между специализированными тегами разметки $и$ (рисунок 4). В качестве инструмента анализа использован язык регулярных выражений. В МВ выделяются: символы арифметических и логических операций, переменные, переменные с индексами, ключевые слова, числа. Если МВ содержит только одну переменную или переменную с индексом, то оно классифицируется как **переменная**. Иначе МВ классифицируется как **формула**.

Рисунок 4. Структура математической формулы

Шаг 2. Связывание формул с переменными. Во всем анализируемом тексте для каждой **переменной** производится поиск вхождения этой переменной в каждую **формулу**. Пусть {F} – множество формул, а {P} – множество переменных. Для $\forall\ p_i \in P$, если $p_i \subset f_k \in F$, устанавливается отношение $\langle p_i, f_k \rangle$. Для каждого отношения в качестве атрибута запоминаются позиции формул и переменных тексте.

В результате разметки получаем отношение много-ко-многим между формулами и переменными, входящими в состав этих формул.

Поиск в Википедии. Система поиска в Википедии включает следующие взаимосвязанные подсистемы: загрузки и анализа данных Википедии, полнотекстового индексирования, индексирования математических формул и переменных, взаимодействия с пользователем, поиска и ранжирования.

Единицей анализируемой и загружаемой информации является html-страница Википедии, содержащая формулы. В системе производится индексирование входных документов как текстовых данных (библиотека Apache Lucene), а также дополнительное позициональное индексирование, построенное в результате применения **метода разметки математических выражений.**

Пользовательский интерфейс реализован как веб-приложение. На странице ввода запроса пользователь может указать в текстовых полях одно или более названий параметров, которые должны присутствовать в искомой формуле. Результаты поиска представляются наборами фрагментов, содержащих нетекстовый объект (формулу), и фрагментов, связывающих переменные формулы и их текстовые определения вне зависимости от их местонахождения в тексте.

На странице результатов поиска (рисунок 5) выводится запрос пользователя и ранжированный по релевантности список ссылок на страницы Википедии.

Поиск и ранжирование выполняются в два этапа. На первом этапе производится полнотекстовый поиск всех вхождений ключевых словосочетаний (**терминов**) в тексты. Для каждого вхождения определяется, существует ли в некоторой окрестности ключевой фразы **переменная**. Для определения окрестности введено понятие максимально допустимого расстояния (МДР) – расстояние в символах влево и вправо от термина, в пределах которого может находиться переменная. По найденным переменным определяется соответствующая формула. Для каждой формулы строится группа текстовых фрагментов, включающих термины и переменные.

На втором этапе производится поиск наилучшей группы текстовых фрагментов для всей совокупности введенных ключевых фраз. Для этого составляются все возможные сочетания полученных текстовых фрагментов в документе и проверяются по критерию близости. В качестве критерия близости использован минимум среднеквадратичного отклонения найденных фрагментов с

определениями переменных от позиции соответствующей формулы. В результате для каждого документа получим оптимальную группу текстовых фрагментов и относящуюся к ним формулу. Результаты для всех документов сортируются по критерию близости и дополнительным критериям релевантности, связанных с полнотой вхождения выделенных переменных в формулу.

Алгоритм поиска математических формул по названию их параметров показал достаточную релевантность в сочетании с высокой скоростью поиска. Проведенное тестирование выявило, что выдаваемые результаты практически всегда имеют непосредственное отношение к задаваемому запросу, и на первой странице поиска находится формула, отвечающая запросам пользователя.

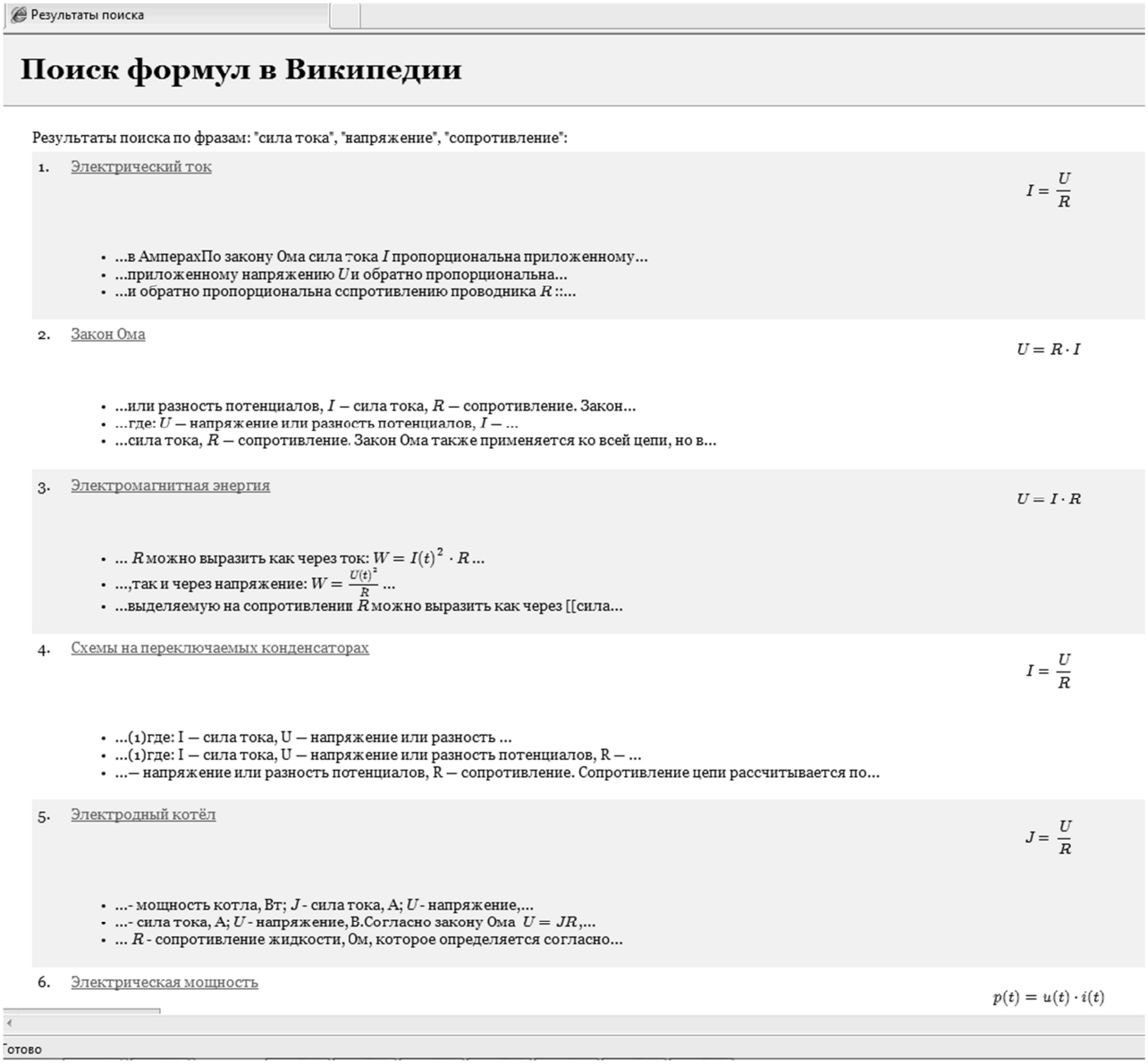

Рисунок 5. Закон Ома. Поиск по параметрам: сила тока, напряжение, сопротивление

Для дальнейшего улучшения релевантности необходимо модифицировать механизм поиска, а не механизм ранжирования. Наиболее очевидным путем представляется синтаксический анализ текста с целью выделения терминологических словосочетаний – именных групп (ИГ) – и дальнейшего анализа отношения выделенных ИГ и заданных пользователем наименований параметров, также являющихся ИГ. С этой целью был разработан подход к разметке для поиска по онтологии с использованием связывания формул с ИГ.

Разметка для поиска по онтологии. Связная коллекция математических документов, размеченная для построения онтологии, дополняется разметкой переменных и формульных выражений, а также их связей с ИГ. Построенная на основе коллекции онтология содержит компоненты онтологии, переменные, формульные выражения, а также связи между ними. Поисковый запрос пользователя на естественном языке переводится в термины онтологии, по которым затем формируется запрос на языке запросов к RDF-документам SPARQL.

Автоматизация построения онтологии упрощает пополнение коллекции математических текстов, что играет важную роль при работе с большими данными. В диссертационной работе рассматривается только дополнение связной коллекции математических текстов, размеченных для построения онтологии, разметкой формул и входящих в них переменных, а также их связей с ИГ.

Модуль формульной разметки реализован на языке Java в формате плагина к текстовому процессору GATE, для разметки используются средства работы с аннотациями библиотеки Gate и оригинальные алгоритмы.

Формульная разметка применяется к XML-документу, предварительно размеченному стандартными аннотациями (Token, Sentence, Math и др.) и NLP-аннотациями (TERM, ENDS). Ключевыми аннотациями для работы алгоритма являются аннотации Math, размечающие формульные фрагменты, и аннотации TERM, соответствующие именным группам.

На основе аннотаций Math строится внутренняя модель документа, содержащая набор разобранных, классифицированных, связанных между собой формульных фрагментов.

Обработка XML-документа включает в себя следующие действия:

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

В первых двух действиях использован **метод разметки математических выражений**, описанный в начале главы. На третьем шаге вводится понятие максимально допустимого расстояния (МДР) между аннотациями Math и TERM, которое определяется как наибольшее расстояние в символах между концом левой аннотации и началом правой, при котором может быть выполнено связывание. МДР является параметром, который оказывает непосред-

ственное влияние на точность связывания и может различаться для разных коллекций документов.

В документах встречается различное взаимное расположение формул и ИГ:

- ИГ содержит формулу. Тогда ИГ – единственный кандидат для связывания. В простейшем случае ИГ состоит из единственного главного слова. В более сложном случае она содержит более одного слова, и рассматривается расстояние между формулой и главным словом. Если это расстояние составляет более трех символов, формула считается дополнением и не связывается;
- Формула и ИГ следуют друг за другом (в пределах одного предложения).

В этом случае основой анализа является концепция МДР. Алгоритм позволяет связывать формулу только с одной именной группой, но с одной и той же ИГ может быть связано более одной формулы.

На заключительном этапе связи, построенные на внутренней модели, переносятся в обрабатываемый документ и используются при построении онтологии связанной коллекции документов.

Проведена оценка **релевантности и полноты связывания** выявленных в тексте математических выражений и именных групп, которая основана на экспертной оценке качества связывания на корпусах математических текстов. В качестве документов для разметки использовались статьи журнала «Известия вузов. Математика» за 1997–2009 гг. Результаты анализа связывания при изменении МДР от 15 до 40 симметрично в обе стороны представлены в таблице 2.

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

Таблица 2. Статистика связывания в зависимости от МДР

Для каждого заданного значения МДР на всем корпусе текстов определялись следующие параметры: **Math** – количество выделенных формул; **Terms** – количество выделенных ИГ; **VirOK** – процент правильных связываний формул с ИГ (полнота связывания); **NotVirOK** – процент правильных несвязываний формул с ИГ (то есть констатация того факта, что математическое выражение находится в контексте, не содержащем его определения); **TotalOk** – общий процент правильно обработанных формул (сумма VirOK и NotVirOK); **VirBad** – процент неправильных связываний формул с ИГ (из возможных кандидатов на связывание была выбрана не подходящая по семантике ИГ или произошло свя-

зывание математического выражения в контексте, не предполагающем связывания); **Others** – другие ошибки связывания (отсутствие связывания там, где оно должно было быть; неправильное выделение ИГ; нераспознанные ИГ; влияющие на связывание особенности оформления текста автором); **TotalBad** – общий процент неправильно обработанных формул (сумма VirBad и Others).

Из таблицы 2 видно, что процент правильно обработанных формул TotalOk и процент ошибок всех типов TotalBad изменяются незначительно, что свидетельствует об устойчивости применяемого алгоритма. Тем не менее, процент правильно связанных математических выражений VirOK имеет тенденцию к возрастанию с увеличением МДР, что вполне ожидаемо. Общий процент ошибок связывания TotalBad также растет с увеличением МДР. Вместе с тем, изменения этих параметров имеют нелинейную зависимость, что позволяет сделать выбор оптимального МДР, при котором отношение VirOK/TotalBad максимально. Для данного корпуса текстов оптимальное МДР составляет 20 символов.

В целом описанные реализации показали принципиальную работоспособность и хорошую устойчивость результатов при применении предложенных подходов, а также выявили ряд проблем с релевантностью поиска. Для повышения релевантности и полноты связывания можно определить следующие направление дальнейшего развития предлагаемого подхода: дополнительный анализ контекста документа; разработка шаблонов, типичных для математических текстов; использование Байесовских методов; обучение распознаванию ключевых слов и конструкций на основе экспертного связывания.

В Заключении представлены основные выводы по работе.

ОСНОВНЫЕ ВЫВОДЫ И РЕЗУЛЬТАТЫ РАБОТЫ

- Разработаны методики: представления структуры РБД в формализме онтологий; построения онтологии предметной области на основе логической модели «сущность–связь», представленной в виде ER-диаграмм, для предметной области, связанной с нефтедобычей; построен лингвистический тезаурус языка этой предметной области;
- Разработаны: 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;
- Создана программная система семантического поиска математических выражений в статьях Википедии;
- Реализована программная система разметки естественнонаучных текстов, содержащих математические выражения, для поиска по онтологии.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях из перечня ВАК

1. Гусенков, А.М. Интеграция реляционных баз данных на основе онтологий. / Е.В. Биряльцев, А.М. Гусенков // Ученые записки Казанского государственного университета. Серия Физико-математические науки. – 2007. – Т. 149, кн. 2. – С. 13–34. – 1,56 п.л. // 0,78 п.л.

2. Гусенков, А.М. Некоторые подходы к разметке естественнонаучных текстов, содержащих математические выражения / Е.В. Биряльцев, А.М. Гусенков, О.Н. Жибрик // Ученые записки Казанского университета. Серия Физико-математические науки. – 2014. – Т. 156, кн. 4. – С. 133–148. – 1,08 п.л. // 0,43 п.л.

Монография

3. Гусенков, А. Интеллектуальный поиск в структурированных массивах информации / Александр Гусенков, Евгений Биряльцев, Ольга Жибрик // LAP LAMBERT Academic Publishing. – Deutschland: OmniScriptum Marceting DEU GmbH, ISBN 978-3-659-76919-1, 2015. – 129 с. – 7,5 п.л. // 3,75 п.л.

Публикации в других изданиях

4. Гусенков, А.М. Интеллектуальный поиск сложных объектов в массивах больших данных / А.М. Гусенков // Электронные библиотеки. – 2016. – Т. 19, № 1. – С. 3–39. – 1,7 п.л.

5. Гусенков, А.М. Поиск математических выражений в естественно-научных текстах. Экспериментальная оценка релевантности / Е.В. Биряльцев, А.М. Гусенков, О.Н. Жибрик // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2014. – Казань: Фэн Академии наук РТ, 2014. – С. 34–37. – 0,27 п.л. // 0,11 п.л.

6. Гусенков, А.М. Некоторые подходы к повышению релевантности поиска математических выражений в естественнонаучных текстах / Е.В. Биряльцев, М.Р. Галимов, А.М. Гусенков, О.Н. Жибрик // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2012. – Казань: Фэн Академии наук РТ, 2012. – С. 78–92. – 0,64 п.л. // 0,22 п.л.

7. Гусенков, А.М. Один подход к реализации нерегламентированного доступа к реляционным базам данных / Е.В. Биряльцев, А.М. Гусенков, С.В. Миронов // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2008. – Казань: Казанский государственный университет, 2009. – С. 10–23. – 0,64 п.л. // 0,25 п.л.

8. Гусенков, А.М. Проблемы управления информационными системами с динамически изменяющейся структурой данных / Е.В. Биряльцев, А.М. Гусенков, С.В. Миронов // Научно-исследовательский институт математики и механики им. Н.Г. Чеботарева Казанского государственного университета. 2003–

2007 гг. – Казань: Казанский государственный университет, 2008. – С. 477–508. – 1,8 п.л. // 0,72 п.л.

9. Гусенков, А.М. Некоторые структурные аспекты интеграции реляционных баз данных / Е.В. Биряльцев, А.М. Гусенков // Труды Третьей международной конференции по когнитивной науке, 20–25 июня 2008 г. – М.: Художественно-издательский центр, 2008. – Т. 2. – С. 562–563. – 0,12 п.л. // 0,06 п.л.

10. Гусенков, А.М. Построение онтологии предметной области на основе логической модели баз данных / Е.В. Биряльцев, А.М. Гусенков // Труды Всероссийской конференции с международным участием «Знания–Онтологии–Теории» (ЗОНТ-07). – Новосибирск: Институт математики им. С.Л. Соболева СО РАН, 2007. – Т. 1. – С. 176–183. – 0,5 п.л. // 0,25 п.л.

11. Гусенков, А.М. Представление структуры реляционных баз данных в формализме онтологий / Е.В. Биряльцев, А.М. Гусенков, Я.Г. Косинов // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2006. – Казань: Отечество, 2007. – С. 32–37. – 0,4 п.л. // 0,16 п.л.

12. Гусенков, А.М. Онтологии реляционных баз данных. Лингвистический аспект / Е.В. Биряльцев, А.М. Гусенков // Труды международной конференции Диалог-2007. – М.: Издательский центр РГГУ, 2007. – С. 50–53. – 0,36 п.л. // 0,18 п.л.

13. Гусенков, А.М. О доступе к электронным коллекциям в виде реляционных баз данных на основе онтологий / Е.В. Биряльцев, А.М. Гусенков, А.М. Елизаров // Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL-2007». – Переславль-Залесский: Ин-т программных систем РАН, 2007. – С. 211–216. – 0,64 п.л. // 0,21 п.л.

14. Гусенков, А.М. Представление модели данных Epicenter POSC на языке онтологий OWL / Е.В. Биряльцев, А.М. Гусенков, А.И. Хайруллина // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2006. – Казань: Отечество, 2007. – С. 38–49.. – 0,48 п.л. // 0,2 п.л.

15. Гусенков, А.М. Особенности лексико-семантической структуры наименований артефактов реляционных баз данных / Е.В. Биряльцев, А.М. Гусенков, М.Р. Галимов // Труды казанской школы по компьютерной и когнитивной лингвистике TEL-2005. – Казань: Казанский государственный университет, 2006. – С. 4–12. – 0,54 п.л. // 0,22 п.л.

16. Гусенков, А.М. Интеграция баз данных на основе онтологий / Е.В. Биряльцев, А.М. Гусенков, В.Д. Соловьев // Инфокоммуникационные технологии глобального информационного общества: Тезисы докладов 3-ей ежегодной международной научно-практической конференции. – Казань: Казанский государственный университет, 2005. – С. 40–41. – 0,12 п.л. // 0,04 п.л.