УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ»

УДК 004.75

БОРОДАЕНКО
Дмитрий Сергеевич

МЕТОД, АЛГОРИТМЫ И ПРОГРАММНАЯ РЕАЛИЗАЦИЯ ОТОБРАЖЕНИЯ РЕЛЯЦИОННЫХ БАЗ ДАННЫХ НА МОДЕЛЬ ДАННЫХ RDF

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

по специальности 05.13.11 - математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Минск 2010

Работа выполнена в учреждении образования "Белорусский государственный университет информатики и радиоэлектроники"

Научный руководитель: **Вишняков Владимир Анатольевич**, доктор технических наук, профессор, заведующий кафедрой менеджмента учреждения образования "Минский институт управления"

Официальные оппоненты: **Голенков Владимир Васильевич**, доктор технических наук, профессор, заведующий кафедрой интеллектуальных информационных технологий учреждения образования "Белорусский государственный университет информатики и радиоэлектроники"

Красько Ольга Владимировна, кандидат технических наук, ведущий научный сотрудник государственного научного учреждения "Объединенный институт проблем информатики Национальной академии наук Беларуси"

Оппонирующая организация: Белорусский государственный университет

Защита состоится 28 апреля 2011 г. в 14 часов на заседании совета по защите диссертаций Д 02.15.04 учреждения образования "Белорусский государственный университет информатики и радиоэлектроники" по адресу: 220013, г. Минск, ул. Бровки, д. 6, тел. 293-89-89, электронный адрес dissovet@bsuir.by.

КРАТКОЕ ВВЕДЕНИЕ

Смена парадигмы распределённого гипертекстового пространства Web, получившая название "Web 2.0", ознаменовала переход от понимания Web как сети документов к сети приложений. Для повышения степени автоматизации доступа к данным в Web консорциум W3C с 1997 года работает над набором технологий Semantic Web, одним из ключевых компонентов которого является язык описания Web-ресурсов Resource Description Framework (RDF).

Среди препятствий на пути массового внедрения RDF в Web наиболее существенные — необходимость переноса уже накопленной информации на модель данных RDF и неудовлетворительная производительность существующих семантических систем при обработке больших объёмов RDF-данных. Наиболее перспективным подходом к решению обеих проблем является интеграция RDF и реляционных данных, позволяющая обеспечить семантический доступ к существующим БД и в то же время использующая испытанные технологии реляционных СУБД для повышения эффективности хранения и обработки RDF-данных. Таким образом, отображение реляционных данных на модель RDF является актуальным направлением исследования.

В отличие от существующих семантических систем, сводящих интеграцию RDF и реляционных СУБД либо к хранению всего набора RDF-данных в таблице триплетов, либо к отображению реляционных данных на RDF, предложенное в данной работе комбинированное решение позволяет совместить преимущества обоих подходов: одновременно достичь скорости доступа на уровне реляционных СУБД для данных, соответствующих реляционной модели, и использовать для остальных RDF-данных таблицу триплетов [13–A].

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Связь работы с крупными научными программами (проектами) и темами

Диссертационное исследование выполнено в рамках следующих НИР:

– "Модели и средства Интернет-маркетинга и интеграции информационных ресурсов для электронного бизнеса предприятий Республики Беларусь" (номер госрегистрации 03-3.1/Мен);

– "Инновационные направления подготовки менеджеров и маркетологов с углубленным изучением информационных технологий" (номер госрегистрации 2006763).

Цель и задачи исследования

Целью диссертационной работы является разработка метода, алгоритмов и программных средств хранения и доступа к семантическим данным на основе отображения реляционных БД. Цель определяет следующие задачи исследования:

1. Анализ методов и средств хранения и семантического доступа к данным, выбор подхода, наиболее эффективного с точки зрения гибкости и скорости доступа.

2. Разработка метода семантического доступа к данным на основе отображения реляционных БД на модель RDF.

3. Разработка модели адаптации реляционных данных для отображения на RDF.

4. Разработка алгоритма преобразования запросов к данным RDF в запросы SQL.

5. Разработка алгоритма обновления реляционных данных по запросу RDF.

6. Реализация программной системы хранения RDF-данных на основе разработанных метода и алгоритмов.

7. Оценка функциональных возможностей и производительности разработанного решения.

Объектом исследования является стандартная модель семантических данных RDF. Предметом исследования является применение отображения реляционных данных на модель RDF для повышения гибкости и скорости доступа к RDF-данным.

Положения, выносимые на защиту

1. Метод семантического доступа к данным на основе отображения реляционных БД на модель RDF, включающий модель адаптации реляционных данных для отображения на RDF, процедуры логического вывода, алгоритм преобразования запросов к данным RDF в запросы SQL, алгоритм обновления реляционных данных по запросу RDF. Разработанный метод отличается от аналогов возможностью комбинировать отображённые реляционные данные с произвольными реифицированными утверждениями RDF, представленными в виде таблицы триплетов.

2. Алгоритм преобразования запросов к данным RDF в запросы SQL, позволяющий при помощи запросов на основе графовых шаблонов производить выборку по базе знаний RDF, сформированной с использованием разработанной модели адаптации реляционных данных. Алгоритм поддерживает следующие выразительные возможности, отсутствующие в аналогичных ре-

шениях: реификация утверждений RDF, автоматическое распознавание вложенных подшаблонов, логический вывод на правилах для подклассов, подотношений и транзитивных отношений на уровне реляционной СУБД.

3. Алгоритм обновления реляционных данных по запросу RDF, предоставляющий отсутствующую в существующих решениях возможность использовать единый язык запросов как для доступа к данным RDF, так и для их обновления, что позволяет сократить время обучения программистов и расходы на разработку ПО.

4. Программная реализация системы хранения RDF-данных на основе разработанного метода и алгоритмов, показавшая на больших наборах данных (до 25 млн. утверждений RDF) скорость обработки запросов по метрике BSBM не менее чем на 40 % выше показателей существующих аналогов. Данная система отличается от известных следующими функциональными особенностями: поддержка СУБД PostgreSQL, MySQL и SQLite; возможность распределения вычислительной нагрузки по преобразованию запросов; реализация логического вывода на уровне СУБД на языке PL/SQL. Поскольку данная разработка построена на основе свободного ПО, она может быть использована в отечественных ИТ-проектах без вовлечения дорогостоящих импортных программных продуктов.

Личный вклад соискателя

Диссертационное исследование является квалификационной научной работой, выполненной соискателем самостоятельно на основе изучения отечественной и иностранной литературы, математических моделей, теорий алгоритмов в области представления и обработки семантических данных.

Основные выводы, теоретические положения и практические разработки принадлежат автору диссертации и составляют содержание данной работы. Научный руководитель принимал участие в постановке задач, определении возможных путей решения, оценке результатов.

Апробация результатов диссертации

Результаты диссертационной работы докладывались и обсуждались на III, IV и VIII республиканских научно-практических конференциях "Управление в социальных и экономических системах" (Минск, 2000–2002), I международной конференции "Информационные системы и технологии" (Минск, 2002), III международной конференции разработчиков систем управления контентом с открытым исходным кодом OSCOM (Бостон, США, 2003), I международной конференции разработчиков и пользователей языка программирования Ruby EuRuKo (Карлсруэ, Германия, 2003), I

международной конференции разработчиков программного обеспечения для сети Интернет BarCamp Sheffield (Шеффилд, Великобритания, 2007), IV международной конференции разработчиков и пользователей свободного программного обеспечения LVEE (Гродно, 2008), I международной конференции по интеллектуальным вычислениям и интеллектуальным системам IEEE ICIS (Шанхай, КНР, 2009).

Опубликованность результатов диссертации

По материалам выполненных исследований опубликовано 14 научных работ, в том числе 3 статьи в рецензируемых изданиях. Общее количество опубликованных материалов составляет 3,63 авторских листов, из них автору принадлежит 3,4 авторских листов. Без соавторства опубликовано 10 работ, из них 2 статьи в рецензируемых изданиях.

Структура и объем диссертации

Диссертация состоит из титульного листа, оглавления, введения, общей характеристики работы, основной части, состоящей из четырёх глав, заключения, библиографического списка и приложений. Работа содержит 88 страниц основного текста, включая 30 рисунков и 4 таблицы; библиографический список, включающий 91 наименование на 8 страницах, а также 6 приложений на 41 странице. В главе 1 рассматривается семантическая модель данных RDF и проблема хранения и доступа к таким данным. В главе 2 разработаны метод, модель и алгоритмы, обеспечивающие семантический доступ к данным на основе отображения реляционных БД на модель RDF. В главе 3 разработана программная реализация системы хранения RDF-данных на основе предложенных алгоритмов и рассмотрены результаты её внедрения в различных приложениях. В главе 4 проведена сравнительная оценка функциональных возможностей и производительности разработанной системы хранения RDF-данных и предложены направления дальнейшей работы.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** приведена область исследования диссертационной работы и обоснована ее актуальность.

В **общей характеристике работы** сформулированы цель и задачи работы, изложены основные положения, выносимые на защиту.

В **первой главе "Анализ методов и средств хранения и доступа к RDF-данным"** выявлен наиболее перспективный подход к хранению RDF-данных: отображение реляционных БД на модель RDF, позволяющее воспользоваться средствами реляционных СУБД для повышения скорости доступа к RDF-данным. Показано, что хранение RDF-данных, которые не

могут быть представлены в реляционной модели, требует сочетания вышеупомянутого подхода с подходом на основе таблицы триплетов.

Определены требования к системе хранения RDF-данных на основе отображения реляционных данных на RDF, такие как: использование языка запросов, предоставляющего более высокий уровень абстракции по сравнению с прикладными программными интерфейсами; прямое обращение к СУБД без посредничества промежуточного API; поддержка выразительных возможности языка RDF-запросов SPARQL; расширяемость языка запросов.

Выявлены ограничения существующих систем, обеспечивающих обработку RDF-данных: использование либо только реляционных данных, либо только таблицы триплетов; низкая скорость обработки запросов; недостаточная широта внедрения хранилищ RDF-данных на основе отображения реляционных данных; слабая поддержка реификации утверждений RDF.

Во **второй главе "Метод и алгоритмы отображения реляционных данных на модель RDF"** разработан метод семантического доступа к данным на основе отображения реляционных БД на модель RDF. Составляющие метода представлены на рисунке 1.

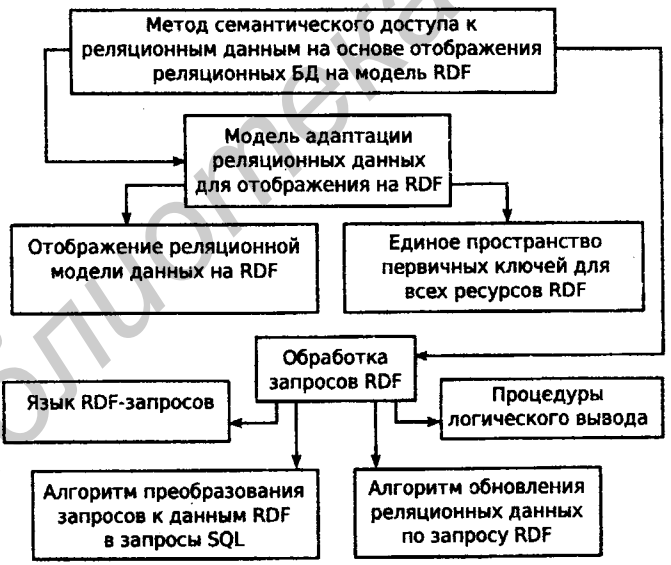


Рисунок 1 – Структура метода семантического доступа к данным на основе отображения реляционных БД на модель RDF

Разработана модель адаптации реляционных данных для отображения на RDF, позволяющая расширить отображённые данные произвольными ре-

ифицируемыми утверждениями RDF, сохранив совместимость с существующими SQL-запросами и обеспечив возможность доступа и обновления данных с использованием языка RDF-запросов.

Процесс адаптации в соответствии с разработанной моделью включает добавление в базу данных атрибутов, внешних ключей, таблиц и хранимых процедур, необходимых для преобразования запросов RDF и поддержки дополнительных возможностей RDF, таких как реификация утверждений и логический вывод. Процесс включает следующие шаги:

1. Ввести n множеств кортежей T_i, представляющих таблицы реляционной базы данных:

$$\begin{aligned} &T_1 = \{\langle a_{t_1 1}, \ldots, a_{t_1 m_1}\rangle\}, \\ &\ldots, \\ &T_n = \{\langle a_{t_n 1}, \ldots, a_{t_n m_n}\rangle\}, \end{aligned} \tag{1}$$

   где $a_{t_1 1}, \ldots, a_{t_n m_n}$ — реляционные атрибуты.

2. Для каждого реляционного атрибута $a_{t_i j}$ выбрать соответствующее отношение $p_k \in P$, где P — множество отношений RDF, построить отображение M отношений RDF на таблицы и атрибуты:

$$M : \{p_k\} \to \{\langle T_i, a_{t_i j}\rangle\}. \tag{2}$$

3. Создать единое пространство первичных ключей как для ресурсов RDF, отображённых из записей таблиц, так и ресурсов, описываемых в таблице триплетов:

   3.1. Создать таблицу ресурсов R, отображённую на суперкласс *rdfs:Resource*, с автоматически генерируемым первичным ключом $\mathrm{id}(R)$, так что для любого определённого на *rdfs:Resource* RDF-отношения p_{Ri}:

$$M(p_{Ri}) = \langle R, a_{Ri}\rangle, \tag{3}$$

   где a_{Ri} — атрибут таблицы R.

   3.2. Заменить первичные ключи $\mathrm{id}(T_i)$ таблиц T_i, отображённых на подклассы класса *rdfs:Resource*, на внешние ключи, ссылающиеся на таблицу ресурсов R, так что:

$$\begin{aligned} &\mathrm{id}(R) = \textstyle\bigcup_{i=1}^{n} \mathrm{id}(T_i), \\ &\forall i \neq j \ \mathrm{id}(T_i) \cap \mathrm{id}(T_j) = \emptyset. \end{aligned} \tag{4}$$

   Обновить существующие внешние ключи с учётом замененных значений первичных ключей.

4. Зарегистрировать хранимые процедуры логического вывода на правилах для *rdfs:subClassOf* для обновления таблицы ресурсов и поддержки целостности внешних ключей при выполнении операций над таблицами подклассов T_i.
5. Создать хранимые процедуры и вспомогательные структуры данных, необходимые для поддержки дополнительных возможностей алгоритма преобразования запросов:
   5.1. Зарегистрировать хранимые процедуры для прочих случаев логического вывода на правилах для *rdfs:subClassOf*.
   5.2. Для представления RDF-данных, не отображённых на реляционную схему, и реификации утверждений RDF создать таблицу триплетов S, отображённую на R так что:

$$\mathrm{id}(S) \subset \mathrm{id}(R)\,. \tag{5}$$

   5.3. Для поддержки логического вывода на правилах для *rdfs:subPropertyOf* добавить атрибуты a_s различения подотношений, ссылающиеся на записи в таблице ресурсов R, хранящие идентификаторы URIref соответствующих отношений, для каждого атрибута $a_p(T_i)$, отображённого на отношение, для которого определены подотношения:

$$a_s : \{\langle \mathrm{id}(T_i), a_p \rangle\} \to P\,. \tag{6}$$

   5.4. Создать таблицы транзитивных замыканий $T^+_{a_t}$ и зарегистрировать хранимые процедуры логического вывода на правилах для *owl:TransitiveProperty* для каждого атрибута $a_t(T_i)$, отображённого на транзитивное отношение p_t, такое, что для любых RDF-ресурсов a, b и c верно:

$$\langle a, p_t, b \rangle \wedge \langle b, p_t, c \rangle \Rightarrow \langle a, p_t, c \rangle\,. \tag{7}$$

В рамках обобщённого алгоритма логического вывода предложен алгоритм обновления транзитивного замыкания, позволяющий реализовать логический вывод на правилах для транзитивных отношений. Перенос логического вывода на уровень СУБД при помощи механизма хранимых процедур позволил выполнять RDF-запросы, опирающиеся на правила логического следования для подклассов, подотношений и транзитивных отношений, за один проход, без свойственных аналогам дополнительных затрат вычислительных ресурсов на последовательное выполнение нескольких вариантов запроса либо на обработку промежуточных результатов на прикладном уровне.

Разработана расширенная версия языка RDF-запросов Squish, обеспечивающая набор возможностей для поиска, извлечения и обновления данных RDF, сопоставимый с возможностями стандартного языка RDF-запросов SPARQL, но более простой в использовании.

Разработаны алгоритм преобразования запросов к данным RDF в запросы SQL и алгоритм обновления реляционных данных по запросу RDF, позволяющие производить выборку и обновление данных по графу RDF, составленному из отображённых реляционных данных и произвольных реифицированных утверждений RDF. В отличие от известных аналогов, разработанный алгоритм преобразования запросов поддерживает как необязательные и отрицательные графовые шаблоны, так и автоматическое распознавание рекурсивных подшаблонов, а также выполнение агрегатных функций и логический вывод на правилах для подмножества словарей RDFS и OWL.

Входными данными для алгоритма преобразования являются:

– набор отображений $M = \langle M_{rel}, M_{attr}, M_{sub}, M_{trans} \rangle$, где $M_{rel} : P \to R$, $M_{attr} : P \to \Phi$, $M_{sub} : P \to S$, $M_{trans} : P \to T$; P — множество отображённых отношений RDF, R — множество реляционных таблиц, Φ — множество реляционных атрибутов, $S \subset P$ — подмножество отношений RDF, для которых заданы подотношения, $T \subset R$ — множество транзитивных замыканий;

– графовый шаблон $\Psi = \langle \Psi_{nodes}, \Psi_{arcs} \rangle = \Pi \cup N \cup \Omega$, где Π, N и Ω — основной, отрицательный и необязательный графовые шаблоны соответственно, такие что Π, N и Ω не имеют общих дуг, и при этом Π, $\Pi \cup N$ и $\Pi \cup \Omega$ образуют связные компоненты графа Ψ;

– глобальное условие фильтрации $F_g \in F$ и локальные условия $F_c : \Psi_{arcs} \to F$, где F — множество всех условий над литералами, выразимых в синтаксисе языка запросов Squish.

Основные шаги алгоритма преобразования запросов представлены в блок-схеме на рисунке 2. На рисунке 3 приведен графовый шаблон Ψ для следующего образца RDF-запроса на языке Squish:

```
SELECT ?msg
WHERE (rdf::predicate ?stmt dc::relation)
      (rdf::subject ?stmt ?msg)
      (rdf::object ?stmt ?tag)
      (dc::date ?stmt ?date)
      (s::rating ?stmt ?rating FILTER ?rating >= :threshold)
EXCEPT (dct::isPartOf ?msg ?parent)
OPTIONAL (dc::language ?msg ?original_lang)
         (s::isTranslationOf ?msg ?translation)
         (dc::language ?translation ?translation_lang)
LITERAL ?original_lang = :lang OR ?translation_lang = :lang
GROUP BY ?msg
ORDER BY max(?date) DESC
```

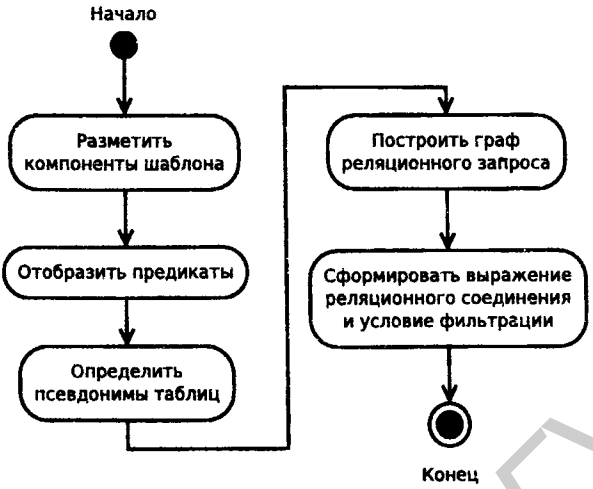
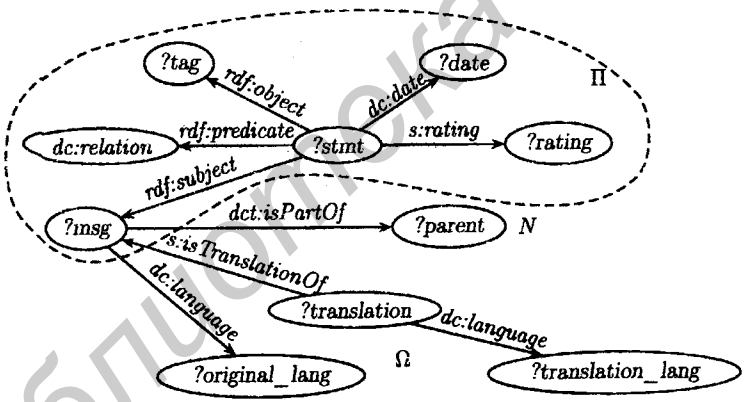


Рисунок 2 – Общая блок-схема алгоритма преобразования запросов к RDF-данным на основе графовых шаблонов в запросы SQL



Pi — обязательный графовый шаблон; *N* — отрицательный графовый шаблон; Ω — необязательный графовый шаблон

Рисунок 3 – Пример графового шаблона Ψ

В результате преобразования данного RDF-запроса будет получен следующий запрос на языке SQL:

```
SELECT DISTINCT a.subject, max(b.published_date)
FROM Statement AS a
INNER JOIN Resource AS b ON (a.id = b.id)
```

```
INNER JOIN Resource AS c ON (a.subject = c.id)
INNER JOIN Message AS d ON (a.subject = d.id)
INNER JOIN Resource AS g ON (a.predicate = g.id)
       AND (g.literal = 'false' AND g.uriref = 'true'
       AND g.label = 'http://purl.org/dc/elements/1.1/relation')
LEFT JOIN (
  SELECT e.language AS _field_b, c.id AS _field_a
  FROM Message AS e
  INNER JOIN Resource AS f ON (f.literal = 'false' AND f.uriref = 'true'
         AND f.label =
             'http://www.nongnu.org/samizdat/rdf/schema#isTranslationOf')
  INNER JOIN Resource AS c ON (c.part_of_subproperty = f.id)
         AND (c.part_of = e.id)
) AS _subquery_a ON (c.id = _subquery_a._field_a)
WHERE (b.published_date IS NOT NULL) AND (a.object IS NOT NULL)
  AND (a.rating IS NOT NULL) AND (c.part_of IS NULL)
  AND (a.rating >= ?) AND (d.language = ? OR _subquery_a._field_b = ?)
GROUP BY a.subject ORDER BY max(b.published_date) DESC
```

В **третьей главе "Программная реализация и использование системы хранения RDF-данных на основе отображения реляционных данных"** определён набор требований к программной системе Graffiti, реализующей разработанные метод и алгоритмы отображения реляционных данных на модель RDF. Разработана архитектура программной системы, предусматривающая поддержку распределённого развёртывания, что обеспечивает более высокую масштабируемость по сравнению с аналогами, реализующими централизованное преобразование запросов. Совместимость архитектуры с широким спектром существующего свободного ПО позволяет использовать данную программную систему в отечественных ИТ-проектах без вовлечения дорогостоящих импортных программных продуктов.

На основе разработанных алгоритмов, требований и архитектуры реализована система хранения и доступа к RDF-данным Graffiti, включающая модуль преобразования запросов, реализованный на языке Ruby и обеспечивающий обработку запросов на языке Squish (рисунок 4), и набор хранимых процедур, реализованный на языке PL/SQL, поддерживающий различные реляционные СУБД и обеспечивающий возможности логического вывода на правилах для подклассов, подотношений и транзитивных отношений над реляционными данными, отображёнными на модель RDF.

Реализована система открытой публикации сообщений в пространстве Web Samizdat, опирающаяся на семантическую модель данных RDF и использующая RDF-хранилище Graffiti в качестве основного средства доступа к данным; использование модели RDF упростило обмен данными с другими приложениями и позволило реализовать полностью прозрачный и децентрализованный процесс структуризации содержания сайтов.

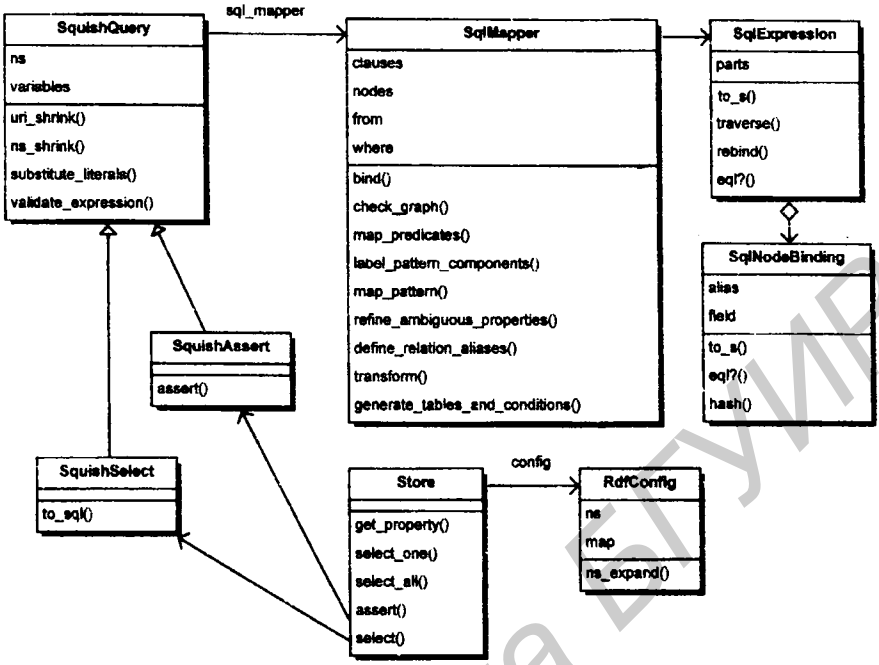


Рисунок 4 – Диаграмма классов разработанной системы хранения RDF-данных

Система хранения RDF-данных Graffiti и система открытой публикации Samizdat внедрены в процесс электронной публикации материалов газеты "Компьютерные Вести", что позволило повысить эффективность модерации материалов, опубликованных пользователями сайта, и соответственно увеличить рентабельность электронной версии газеты за счёт сокращения трудоёмкости информационной поддержки в пересчёте на количество посещений.

Реализован и внедрён в систему поддержки принятия решений ИЧП "Самсолюшнс" модуль семантического поиска и анализа корпоративной информации на основе RDF-хранилища Graffiti и модуля поиска системы открытой публикации Samizdat. Модуль обеспечил руководству компании оперативный доступ к данным различных приложений и позволил учитывать актуальные и точные данные в процессе принятия решений без траты рабочего времени сотрудников на сбор информации из разрозненных приложений.

Результаты работы внедрены в учебный процесс кафедры менеджмента Минского института управления, что позволило перевести процесс обучения по дисциплинам "Интеллектуальные системы в экономике" и "Интеллектуальные системы в управлении" на качественно более высокий уровень

благодаря формированию у обучаемых системного восприятия учебного материала и получению практических навыков в области моделирования семантических структур RDF.

В **четвёртой главе "Оценка функциональных возможностей и производительности системы Graffiti "** показано, что разработанная система хранения RDF-данных Graffiti по своим функциональным возможностям превосходит большинство существующих систем отображения реляционных данных на RDF (таблица 1).

Таблица 1 – Возможности языка запросов в Graffiti и других системах, реализующих динамическое преобразование запросов RDF в SQL

Выразительное средство языка запросов	Система хранения RDF-данных			
	Graffiti	Federate	D2RQ	Virtuoso
Синтаксис SPARQL	нет	нет	да	да
Простой графовый шаблон	да	да	да	да
Условия фильтрации	да	нет	да	да
Необязательный графовый шаблон	да	да	частично[1]	да
Вложенные группы графовых шаблонов	частично[2]	нет	нет	да
Неопределённые предикаты	нет	нет	нет	да
Реификация утверждений	да	нет	нет	нет[3]
Логический вывод для RDFS	частично[4]	нет	нет	частично[5]
Логический вывод для OWL	частично[6]	нет	нет	частично[7]
Агрегация результатов	да	нет	нет	да
Язык обновления данных	да	нет	нет	да

Примечания

[1]поддержка отдельных случаев добавлена в версии 0.6 в 2009 году

[2]необязательные подшаблоны автоматически делятся на связные группы

[3]поддерживается интеграция с таблицей триплетов

[4]вывод для *rdfs:subClassOf* и *rdfs:subPropertyOf* с использованием хранимых процедур

[5]вывод для *rdfs:subClassOf* и *rdfs:subPropertyOf* посредством перебора вариантов

[6]вывод для *owl:TransitiveProperty* с использованием хранимых процедур

[7]вывод для *owl:sameAs* посредством перебора вариантов

Преимущества перед самым развитым из существующих аналогов, системой Virtuoso RDF Views, включают более высокую гибкость (обеспеченную поддержкой реификации утверждений RDF и компактной модульной архитектурой, упрощающей интеграцию в существующие системы) и масштабируемость (достигаемую за счёт использования хранимых процедур логического вывода и возможности распределения вычислительной нагрузки по преобразованию RDF-запросов).

Тестирование системы Graffiti по метрике BSBM показало, что разработанная система превосходит аналоги по скорости обработки запросов и

масштабируемости как в целом (рисунок 5), так и на большинстве видов запросов данной метрики, рассмотренных по отдельности.

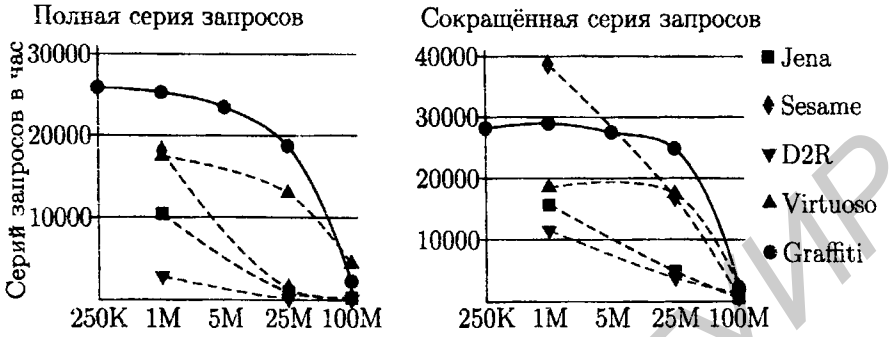


Рисунок 5 – Скорость обработки типовой серии запросов метрики BSBM

По совокупному показателю QMpH (количество обработанных серий запросов в час) разница с самой производительной из аналогичных систем, Virtuoso, составила 45 % при обработке набора данных, насчитывающего 1 млн. триплетов, и 44 % при обработке набора данных из 25 млн. триплетов.

Анализ скорости обработки отдельных запросов метрики BSBM выявил, что более высокая по сравнению с аналогами производительность системы Graffiti обеспечивается следующими особенностями разработанного набора алгоритмов преобразования RDF-запросов и выбранной архитектуры программной реализации: использование хранимых процедур для реализации логического вывода над транзитивными свойствами, оптимизация частного случая вырожденных поддеревьев графа реляционного запроса, перенос свойств суперкласса *rdfs:Resource* в отдельную таблицу, использование стороннего свободного ПО для решения смежных задач (в частности, средств полнотекстового индексирования СУБД PostgreSQL).

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Проанализированы методы и средства хранения и доступа к данным, представленным в семантической модели RDF. Выявлен как наиболее перспективный подход к хранению RDF-данных на основе отображения реляционных данных на модель RDF; показана необходимость сочетания данного подхода с подходом на основе записи RDF-утверждений в таблице трипле-

тов для хранения RDF-данных, которые не могут быть представлены в реляционной модели. Определены следующие требования к системе хранения RDF-данных: использование языка запросов, предоставляющего более высокий уровень абстракции по сравнению с прикладными программными интерфейсами; прямое обращение к СУБД без посредничества промежуточного API; поддержка выразительных возможности языка RDF-запросов SPARQL; расширяемость языка запросов. Выявлены ограничения существующих семантических систем: использование либо только реляционных данных, либо только таблицы триплетов; низкая скорость обработки запросов; недостаточная широта внедрения систем на основе отображения реляционных данных. Результаты изложены в первой главе, опубликованы в следующих работах автора [4–A, 7–A, 8–A, 9–A, 14–A].

2. Разработан метод семантического доступа к данным на основе отображения реляционных БД на модель RDF, включающий модель адаптации реляционных данных для отображения на RDF, процедуры логического вывода, алгоритм преобразования запросов к данным RDF в запросы SQL, алгоритм обновления реляционных данных по запросу RDF. Разработанный метод отличается от аналогов возможностью комбинировать отображённые реляционные данные с произвольными реифицированными утверждениями RDF, представленными в виде таблицы триплетов. Разработана расширенная версия языка RDF-запросов Squish, обеспечивающая набор возможностей для поиска, извлечения и обновления данных RDF, сопоставимый с возможностями стандартного языка RDF-запросов SPARQL, но более простой в использовании. Результаты изложены во второй главе, опубликованы в следующих работах автора [1–A, 5–A, 6–A, 12–A, 13–A].

3. Разработан алгоритм преобразования запросов к данным RDF в запросы SQL, позволяющий производить выборку данных по графу RDF, составленному из отображённых реляционных данных и произвольных реифицированных утверждений RDF. В отличие от известных аналогов, разработанный алгоритм поддерживает как необязательные и отрицательные графовые шаблоны, так и автоматическое распознавание рекурсивных подшаблонов, а также выполнение агрегатных функций и логический вывод на правилах для подмножества словарей RDFS и OWL. Разработан алгоритм обновления реляционных данных по запросу RDF, обеспечивающий отсутствующую в аналогах возможность использования единого языка запросов как для доступа к данным RDF, так и для их обновления. Предложен набор хранимых процедур, позволяющих реализовать правила логического следования для предикатов подклассов, подотношений и транзитивных отношений с более высокой производительностью по сравнению с существующими ана-

логами. Результаты изложены во второй главе, опубликованы в следующих работах автора [1–A, 5–A, 11–A].

4. Разработанные метод и алгоритмы реализованы в программной системе хранения и доступа к RDF-данным Graffiti, включающей модуль преобразования запросов, реализованный на языке Ruby и обеспечивающий обработку запросов на языке Squish, и набор хранимых процедур логического вывода, реализованный на языке PL/SQL. Архитектура системы Graffiti предусматривает поддержку распределённого развёртывания, что обеспечивает более высокую масштабируемость по сравнению с аналогами, реализующими централизованное преобразование запросов. Разработанная система построена на основе свободного ПО, что позволяет использовать её в отечественных ИТ-проектах без вовлечения дорогостоящих импортных программных продуктов. Результаты изложены в третьей главе, опубликованы в следующих работах автора [2–A, 10–A, 14–A].

5. Проведено сравнение функциональных возможностей разработанной системы с существующими аналогами. Выявлен ряд преимуществ: поддержка реификации утверждений RDF, использование хранимых процедур для логического вывода на правилах RDFS и OWL на уровне реляционной СУБД, возможность распределения вычислительной нагрузки по преобразованию RDF-запросов. Производительность разработанной системы измерена по метрике BSBM, полученные показатели скорости обработки запросов превосходят аналогичные системы не менее чем на 40 %. Выявлены следующие причины более высокой производительности: использование хранимых процедур для реализации логического вывода, оптимизация отдельных частных случаев при преобразовании запросов, особенности модели адаптации реляционных данных, использование существующего свободного ПО для решения смежных задач. Результаты изложены в четвёртой главе, опубликованы в следующей работе автора [2–A].

Рекомендации по практическому использованию результатов

1. Программная система хранения RDF-данных Graffiti, реализующая разработанные метод семантического доступа к данным, модель адаптации реляционных данных, хранимые процедуры логического вывода, алгоритмы преобразования запросов и обновления данных, может быть использована как для интеграции существующих реляционных БД посредством технологии RDF, так и для построения новых семантических приложений. Одним из важных практических результатов данной работы является то, что программная

система Graffiti вносит белорусский вклад в решение глобальной проблемы развития Интернета — построения семантического Web-пространства.

2. На основе программной системы Graffiti реализован модуль семантического поиска и анализа корпоративной информации, интегрирующий данные различных приложений и используемый в системе поддержки принятия решений в ИЧП "САМСОЛЮШНС" (акт о практическом использовании от 03.08.2010).

3. Реализована программная система открытой электронной публикации сообщений Samizdat, опирающаяся на семантическую модель данных RDF и использующая систему хранения RDF-данных Graffiti в качестве основного средства доступа к данным. Использование модели RDF упростило обмен данными с другими приложениями и позволило реализовать полностью прозрачный и децентрализованный процесс структуризации содержания сайтов. Программная система Samizdat использована в процессе электронной публикации материалов газеты "Компьютерные Вести" (акт о практическом использовании от 02.09.2010).

4. Результаты исследования внедрены в учебный процесс Минского института управления и использованы в дисциплинах "Интеллектуальные системы в экономике" и "Интеллектуальные системы в управлении" для специальностей "Менеджмент" и "Маркетинг" (акт о практическом использовании от 26.11.2010).

5. Использование разработанной программной системы в производстве позволило продемонстрировать, как технологии Semantic Web могут быть встроены в существующие приложения, обслуживающие тысячи пользователей, без вовлечения дополнительных аппаратных ресурсов. Так, тестирование скорости обработки запросов программной системы Graffiti на синтетической метрике BSBM показало производительность на 40 % выше, чем в существующих аналогах, а тестирование на наборе данных, используемом в производстве, показало скорость обработки запросов, сопоставимую с производительностью реляционных СУБД. Более высокая гибкость системы Graffiti по сравнению с аналогами позволяет использовать данную систему в более широком спектре приложений и сократить расходы на модификацию информационных систем при изменении требований.

6. Поскольку данная разработка построена исключительно с использованием свободных программных средств, она может быть использована в отечественных ИТ-проектах без вовлечения дорогостоящих импортных программных продуктов.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

Статьи:

1–А. Бородаенко, Д.С. Отображение реляционных данных на семантическую модель RDF при помощи динамического преобразования запросов / Д.С. Бородаенко // Доклады БГУИР. – 2010. – № 2(48). – С. 84-89.

2–А. Бородаенко, Д.С. Функциональные особенности и производительность системы хранения RDF-данных в реляционных СУБД Graffiti / Д.С. Бородаенко // Доклады БГУИР. – 2010. – № 4(50). – С. 95-99.

3–А. Бородаенко, Д.С. Перспективы использования технологии RDF в сети Вуза / Д.С. Бородаенко, В.А. Вишняков // Информатизация образования. – 2010. – № 4. – С. 44-53.

Материалы конференций:

4–А. Бородаенко, Д.С. Автоматизация создания специализированных баз данных на основе поиска в пространстве Web / Д.С. Бородаенко, В.А. Вишняков // Управление в социальных и экономических системах : материалы 8-й междунар. науч.-практ. конф., Минск, 18-19 декабря 2002 г. / МИУ; под. ред. И.П. Суши [и др.]. – Мн., 2002. – С. 115-117.

5–А. Borodaenko, D. On-demand RDF to Relational Query Translation in Samizdat RDF Store / D. Borodaenko // Intelligent Computing and Intelligent Systems : proceedings of the int. conf., Shanghai, November 2009 : in 4 v. / IEEE; ed.: Wen Chen [et al.]. – IEEE Press, 2009. – V. 3. – P. 413-417.

Тезисы докладов:

6–А. Бородаенко, Д.С. Элементы языка системы логического вывода / Д.С. Бородаенко // Управление в социальных и экономических системах : материалы 3-й респ. науч.-практ. конф., Минск, 14-15 марта 2000 г. : в 2 т. / МИУ; под. ред. И.П. Суши [и др.]. – Мн., 2000. – Т. 2. – С. 120.

7–А. Бородаенко, Д.С. Выделение релевантной информации для повышения эффективности работы в WWW / Д.С. Бородаенко // Управление в социальных и экономических системах : материалы 4-й респ. науч.-практ. конф., Минск, 2000 : в 2 т. / МИУ; под. ред. И.П. Суши [и др.]. – Мн., 2001. – Т. 2. – С. 194.

8–А. Бородаенко, Д.С. Матричное кодирование и обработка web-документов / Д.С. Бородаенко, В.А. Вишняков // Современные средства связи : материалы 6-й междунар. науч.-практ. конф., Нарочь, 1-5 октября 2001 г. / Известия Белорусской инженерной академии. – 2001. – № 1(11)/2. – С. 80.

9–А. Бородаенко, Д.С. Проблема организации и обработки данных в Web / Д.С. Бородаенко, В.А. Вишняков // Современные средства связи : материалы 6-й междунар. науч.-практ. конф., Нарочь, 1-5 октября 2001 г. / Известия Белорусской инженерной академии. – 2001. – № 1(11)/2. – С. 80-81.

10–А. Borodaenko, D. RDF model for an open publishing and cooperation engine / D. Borodaenko // Open Source Content Management (OSCOM3) [Electronic resource]: proceedings of int. conf., Boston, MA, USA, May 28-30, 2003 / Harvard University; ed.: Gregor J. Rothfuss [et al.]. – 2003. – Mode of access: http://samizdat.nongnu.org/slides/oscom3_samizdat.html. – Date of access: 04.07.2010.

11–А. Borodaenko, D. RDF storage for Ruby: the case of Samizdat / D. Borodaenko // European Ruby Conference (EuRuKo 2003) [Electronic resource]: proceedings of int. conf., Karlsruhe, Germany, June 21-22, 2003 / University of Karlsruhe. – 2003. – Mode of access: http://samizdat.nongnu.org/slides/euruko2003_samizdat.html. – Date of access: 04.07.2010.

Прочие публикации:

12–А. Borodaenko, D. Samizdat RDF Implementation Report / Borodaenko, D. // RDF Interest ML [Electronic resource]. – W3C, September 2003. – Mode of access:
http://lists.w3.org/Archives/Public/www-rdf-interest/2003Sep/0043.html. – Date of access: 01.08.2010.

13–А. Borodaenko, D. Accessing Relational Data with RDF Queries and Assertions / D. Borodaenko // Samizdat Project [Electronic resource]. – Minsk, April 2004. – Mode of access: http://samizdat.nongnu.org/papers/rel-rdf.pdf. – Date of access: 04.07.2010.

14–А. Borodaenko, D. Model for Collaborative Decision Making Based on RDF Reification / D. Borodaenko // Samizdat Project [Electronic resource]. – Minsk, April 2004. – Mode of access:
http://samizdat.nongnu.org/papers/collreif.pdf. – Date of access: 04.07.2010.

РЭЗЮМЭ

Барадаенка Дзмітрый Сяргеевіч

Метад, алгарытмы і праграмная рэалізацыя адлюстравання рэляцыйных баз дадзеных на мадэль дадзеных RDF

Ключавыя словы: мадэль дадзеных RDF, рэляцыйныя СКБД, мова запытаў, адлюстраванне запытаў, метрыка прадукцыйнасці BSBM.

Мэтай дысертацыйнай работы з'яўляецца распрацоўка метада, алгарытмаў і праграмных сродкаў захоўвання і доступу да семантычных дадзеных на аснове адлюстравання рэляцыйных баз дадзеных.

Аб'ектам даследвання з'яўляецца стандартная мадэль семантычных дадзеных RDF. Прадметам даследвання з'яўляецца прымяненне адлюстравання рэляцыйных дадзеных на мадэль RDF для павышення гібкасці і хуткасці доступу да RDF-дадзеных.

Распрацаваны метад семантычнага доступу да дадзеных на аснове адлюстравання рэляцыйных БД на мадэль RDF; мадэль адаптацыі реляцыйных дадзеных для адлюстравання на RDF; алгарытмы пераўтварэння запытаў да дадзеных RDF у запыты SQL і аднаўлення рэляцыйных дадзеных па запыце RDF; праграмная рэалізацыя сістэмы захоўвання RDF-дадзеных на аснове распрацаваных метада і алгарытмаў.

Навуковая навізна атрыманых вынікаў заключаецца ў спалучэнні пераваг рэляцыйных СКБД і семантычнай мадэлі дадзеных RDF, забеспячэнні магчымасці расшырэння існуючых СКБД семантычнымі дадзенымі, і рэалізацыі адсутнічаючых у існуючых аналагах магчымасцей рэіфікацыі сцвярджэнняў RDF, аўтаматычнага распазнавання ўкладзеных падшаблёнаў, лагічнага вываду на ўзроўні рэляцыйнай СКБД.

Распрацаваная сістэма захоўвання RDF-дадзеных паказала на наборах дадзеных, налічваючых да 25 млн. сцвярджэнняў, прадукцыйнасць апрацоўкі запытаў па мэтрыцы BSBM на 40 і болей адсоткаў вышэй за прадукцыйнасць існуючых аналагаў.

Вынікі работы выкарыстаныя ў працэсе мадэрацыі матэрыялаў сайта газэты "Кампутарные Весткі", у працэсе прымання рашэнняў ЗПП "Самсалюшнс", у навучальным працэсе кафедры мэнэджмэнту Мінскага інстытуту кіравання.

РЕЗЮМЕ

Бородаенко Дмитрий Сергеевич

Метод, алгоритмы и программная реализация отображения реляционных баз данных на модель данных RDF

Ключевые слова: модель данных RDF, реляционные СУБД, язык запросов, преобразование запросов, метрика производительности BSBM.

Целью диссертационной работы является разработка метода, алгоритмов и программных средств хранения и доступа к семантическим данным на основе отображения реляционных баз данных.

Объектом исследования является стандартная модель семантических данных RDF. Предметом исследования является применение отображения реляционных данных на модель RDF для повышения гибкости и скорости доступа к RDF-данным.

Разработаны метод семантического доступа к данным на основе отображения реляционных БД на модель RDF; модель адаптации реляционных данных для отображения на RDF; алгоритмы преобразования запросов к данным RDF в запросы SQL и обновления реляционных данных по запросу RDF; программная реализация системы хранения RDF-данных на основе разработанных метода и алгоритмов.

Научная новизна полученных результатов заключается в сочетании преимуществ реляционных СУБД и семантической модели данных RDF, обеспечении возможности расширения существующих СУБД семантическими данными, и реализации отсутствующих в существующих аналогах возможностей реификации утверждений RDF, автоматического распознавания вложенных подшаблонов, логического вывода на уровне реляционной СУБД.

Разработанная система хранения RDF-данных показала на наборах данных, насчитывающих до 25 млн. утверждений, скорость обработки запросов по метрике BSBM на 40 % и более выше производительности существующих аналогов.

Результаты работы использованы в процессе модерации материалов сайта газеты "Компьютерные Вести", в процессе принятия решений ИЧП "Самсолюшнс", в учебном процессе кафедры менеджмента Минского института управления.

RESUME

Dmitry S. Borodaenko

Method, algorithms, and software implementation for mapping relational databases to the RDF data model

Keywords: RDF data model, relational DBMS, query language, query translation, BSBM benchmark.

The goal of the dissertation is development of a method, algorithms, and software implementation of a semantic data store based on mapping of relational databases.

The object of the research is the RDF standard semantic data model. The subject of the research is the application of relational to RDF data mapping for improving flexibility and speed of RDF data access.

The results of the research include a method for semantic data access based on mapping of relational databases to the RDF data model; model for RDF mapping adaptation of relational data; algorithms for translation of RDF data queries into SQL queries and processing updates of relational data based on RDF queries; software implementation of an RDF store based on the produced method and algorithms.

The following aspects of the achieved results contribute to the scientific novelty of this work: combination of the advantages of relational DBMSes and the RDF semantic data model, ability to extend existing relational databases with semantic data, and implementation of query capabilities missing from the similar systems, such as RDF statement reification, automatic detection of nested sub-patterns, logical inference on the relational DBMS level.

RDF query processing performance of the implemented RDF store was measured using the BSBM benchmark on datasets of up to 25 million statements and exceeded performance of similar systems by 40 and more percent.

Results of the research were successfully used by "Kompyuternye Vesti" newspaper in the site postings moderation system, by FPE "SaM Solutions" in decision making process, by the Minsk Management Institute in the education process.

Научное издание

БОРОДАЕНКО Дмитрий Сергеевич

МЕТОД, АЛГОРИТМЫ И ПРОГРАММНАЯ РЕАЛИЗАЦИЯ ОТОБРАЖЕНИЯ РЕЛЯЦИОННЫХ БАЗ ДАННЫХ НА МОДЕЛЬ ДАННЫХ RDF

специальность 05.13.11 - математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Подписано в печать 21.03.2011.	Формат 60x84 $^{1}/_{16}$.	Бумага офсетная.
Гарнитура «Таймс».	Отпечатано на ризографе.	Усл. печ. л. 1,63.
Уч.-изд. л. 1,4.	Тираж 60 экз.	Заказ 136.