УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО УНИВЕРСИТЕТА
Том 154, кн. 3 Физико-математические науки *2012*

УДК 004.912+004.021

ПРОТОТИП ПРОГРАММНОЙ ПЛАТФОРМЫ ДЛЯ ПУБЛИКАЦИИ СЕМАНТИЧЕСКИХ ДАННЫХ ИЗ МАТЕМАТИЧЕСКИХ НАУЧНЫХ КОЛЛЕКЦИЙ В ОБЛАКЕ LOD

О.А. Невзорова, Н.Г. Жильцов, Д.А. Заикин, О.Н. Жибрик, А.В. Кириллович, В.Н. Невзоров, Е.В. Бирюльцев

Аннотация

В статье представлен прототип программной платформы для извлечения связанных открытых данных (Linked Open Data) из заданной коллекции математических научных статей. Задача получения семантического представления из коллекций выбранной предметной области является актуальной, так как в настоящий момент в облаке связанных открытых данных отсутствуют новейшие данные по профессиональной математике. По нашему мнению, одной из основных причин этого является отсутствие специализированных инструментов, способных анализировать семантику, лежащую в основе статей по математике, и эффективно строить их консолидированное представление. В статье описан комплексный подход к анализу математических документов для представления содержимого статей и их метаданных в формате RDF. Рассмотрены методы и технологии извлечения семантических данных из математических статей на основе специальных онтологий, а также описаны эксперименты по интеграции построенного математического RDF-набора с существующими наборами данных в Интернете.

Ключевые слова: индексация, связанные данные, проектирование онтологий.

Введение

Проект Linked Open Data[1] (LOD) выявил важные преимущества представления объектов гетерогенных данных, полученных от различных контент-провайдеров, в виде единого связанного облака. Главное преимущество заключается в стандартизованном подходе к структурированию и хранению интегрированных данных. Как правило, данные загружаются и представляются в виде RDF (Resource Description Framework[2]), то есть триплетов вида «субъект – предикат – объект», из таких традиционных хранилищ, как реляционные базы данных, либо, реже, из веб-страниц или полуструктурированных текстовых документов, используя так называемые принципы связанных данных, предложенные Т. Бернерсом-Ли[3]. Специализированные схемы данных – онтологии – разрабатываются для моделирования предметных областей. Современные приложения семантического поиска, например, такие как система семантического поиска Sindice[4] или мэшап Sig.ma[5], используют опубликованные RDF-данные для более точной обработки поисковых запросов либо для сбора и отображения информации о сущностях, которые представляют интерес для пользователя.

[1] http://lod2.eu
[2] http://www.w3.org/RDF/
[3] http://www.w3.org/DesignIssues/LinkedData.html
[4] http://sindice.com/
[5] http://sig.ma/

В сентябре 2011 г. объем данных в LOD составлял более 30 млрд. триплетов, хранящихся в примерно 300 наборах данных. Среди доминирующих предметных областей – данные из правительственных источников (43% по числу триплетов), географические данные (22%) и науки о жизни (биология, биохимия, генетика и др. – 9%).

Научные коллекции представлены в LOD в виде неофициальных и не поддерживаемых на постоянной основе наборов данных (ACM[6], DBLP[7] и CiteSeer[8]). Основным содержимым этих наборов данных являются стандартные метаданные статей (название, год публикации, информация об авторах и др.).

Для Казанского университета крайне актуальной является подготовка наборов связанных данных на основе статей научных журналов, издаваемых в университете, для публикации связанных открытых данных в Интернете. В статье рассматривается задача подготовки набора связанных данных для журнала «Известия вузов. Математика» за период с 1997 по 2009 гг. Для подготовки набора связанных данных использовались математические статьи в формате LaTeX. Можно ожидать, что публикация в LOD данных математической коллекции поможет профессиональным исследователям и студентам получить удобный доступ к специальным знаниям, интегрированным в глобальную систему знаний.

В работе рассматриваются вопросы проектирования и реализации программных решений, обеспечивающих извлечение связанных данных для коллекции математических статей. В основе подхода лежит представление коллекции математических документов в виде единого семантического графа, в котором как вершины – объекты математического знания, так и ребра – связи между ними – определяются множеством специализированных словарей (онтологий), которые специфицируют как термины из области математики, так и элементы логической структуры математического документа.

Основными функциями разработанного программного прототипа для публикации данных в LOD являются:

- индексирование математических статей в формате LaTeXв виде LOD-совместимых RDF-данных;
- извлечение метаданных статьи в виде концептов онтологии AKT Portal Ontology[9];
- извлечение логической структуры документа с использованием онтологии Mocassin;
- извлечение экземпляров математических сущностей в виде концептов онтологии OntoMath и связывание с ресурсами DBPedia;
- распознавание семантики формул через связывание полученных экземпляров математических сущностей с математическими выражениями и формулами в тексте;
- установление взаимосвязи между опубликованными RDF-данными и существующими наборами данных LOD.

1. Обзор близких работ

В настоящее время генерация RDF-наборов данных из полуструктурированных или слабоструктурированных документов является актуальной задачей. Существующие программные инструменты в основном фокусируются на непосредственном преобразовании данных и практически не добавляют семантику к данным. Например, библиотека Any23 принимает на вход веб-страницы XHTML, которые

[6]http://acm.rkbexplorer.com/
[7]http://dblp. rkbexplorer.com/
[8]http://citeseer.rkbexplorer.com/
[9]http://www.aktors.org/ontology/

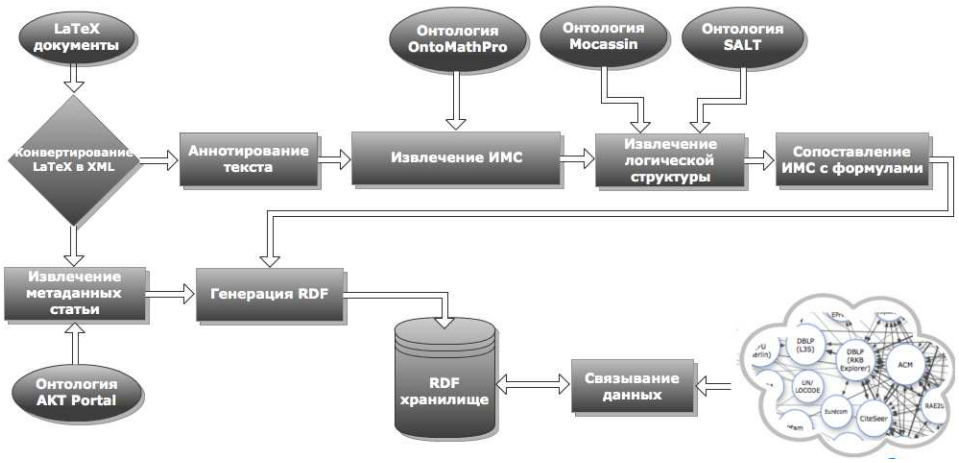

Рис. 1. Архитектура прототипа

могут содержать RDFa-аннотации, и конвертирует их в формат RDF. Эта библиотека недавно подтвердила свою надежность, будучи использованной в процессе масштабного извлечения структурированных данных из открытого хранилища Common Crawl [1].

Представление математических знаний является новой, быстро развивающейся предметной областью. Формальные языки для описания математических сущностей, такие как MathLang [2] и OMDoc [3], позволяют создавать семантические модели математических документов и изначально поддерживают логические структуры на уровне высказываний. Однако создание строго формализованных математических документов является трудоемким процессом. В работе [4] предложен подход, заключающийся в представлении записей лекций по математике в виде связанных данных путем редактирования их специальным макро-пакетом sTeX языка LaTeX. Эта работа в основном сфокусирована на математических формулах и элементах логической структуры и является первой работой, пытающейся связать математические тексты и LOD.

Можно отметить также ряд важных работ из области извлечения онтологий (ontology extraction), в которых решаются задачи анализа текстов в терминах, определенных предметно-ориентированных словарей, таксономий или онтологий. Качество извлечения онтологий достигло впечатляющих результатов в таких областях, как биоинформатика [5], экология [6], юриспруденция [7] и электронная коммерция [8].

Основные задачи, решение которых обсуждается в настоящей статье, были сформулированы в [9]. Здесь представлены результаты, полученные при решении задачи автоматического извлечения RDF-представлений не только метаданных, но и содержания научных статей.

2. Архитектура программного прототипа

Общая инфраструктура процесса опубликования математических данных в LOD приведена на рис. 1.

Прототип состоит из 8 модулей, которые могут быть сгруппированы в следующие подсистемы:

- преобразование формата;
- аннотирование текста;

- семантическое аннотирование;
- аннотирование метаданных;
- генерация RDF;
- связывание.

Следующие подразделы содержат детальное описание каждого из указанных модулей.

2.1. Преобразование формата. В настоящий момент прототип поддерживает только LaTeX-формат входного документа. Можно указать две основные причины подобного решения: во-первых, это связано с необходимостью работы с логической структурой документа и формулами, во-вторых, данный формат имеет высокую популярность среди математиков. Однако работа с данным форматом программно не очень удобна, поэтому в прототипе используется набор инструментов ArXMLiv [10], который включает конвертор LaTeXML[10], используемый для преобразования форматов из LaTeXв XML. ArXMLiv предоставляет мэппинг-файлы для LaTeXML, которые обеспечивают преобразование между командными средами пакетов LaTeXи элементами XML-представлений. Дополнительным преимуществом является то, что в ArXMLiv математические выражения хранятся как отдельные XML-элементы, а их исходные LaTeX-представления – как атрибуты (в настоящей статье будем называть такое представление математических выражений «математическими аннотациями»).

2.2. Аннотирование текста. Подсистема аннотирования текста в ее текущем состоянии поддерживает только обработку документов на русском языке и последовательно решает следующие стандартные лингвистические задачи: токенизация, разделение предложений, морфологический анализ и извлечение именных групп. При токенизации производится извлечение объектов текста и их классификация. Набор типов объектов текста включает следующие типы: Словарное слово, Аббревиатура, Число, Формула, Знак препинания, Сложное слово, Имя собственное, Омоним, Неизвестное слово и др.

После токенизации производится сегментация предложений, однако разрешение многозначности внутренних точек производится в ряде случаев при токенизации. Например, выделение внутренних точек в фамильно-именных группах (например, И.В. Иванов) управляется специальными правилами, базирующимися на регулярных выражениях. Методы сегментации предложений применяются для определения границ предложений на основе анализа знаков препинания, отдельно сегментируются границы зависимых предложений, которые используются при анализе формульных контекстов.

Морфологический анализ базируется на словаре русской грамматики, расширенном специальными терминологическими лексиконами и списками аббревиатур. Результатом анализа является грамматическая аннотация на уровне слов. При этом омонимы аннотируются набором грамматических аннотаций.

В русском языке именная группа (NP) обычно состоит из главного слова, которое обозначается как NP.Head и его левых и правых модификаторов (NP.Dependent). Отношение между главным словом и зависимыми словами является синтаксическим. В лингвистической теории выделяются несколько видов отношений между главным и зависимыми словами словосочетания (именной группы). Построение именных групп осуществляется по правилам, которые опираются на определенную внутреннюю структуру группы (рис. 2).

[10]http://dlmf.nist.gov/LaTeXML/

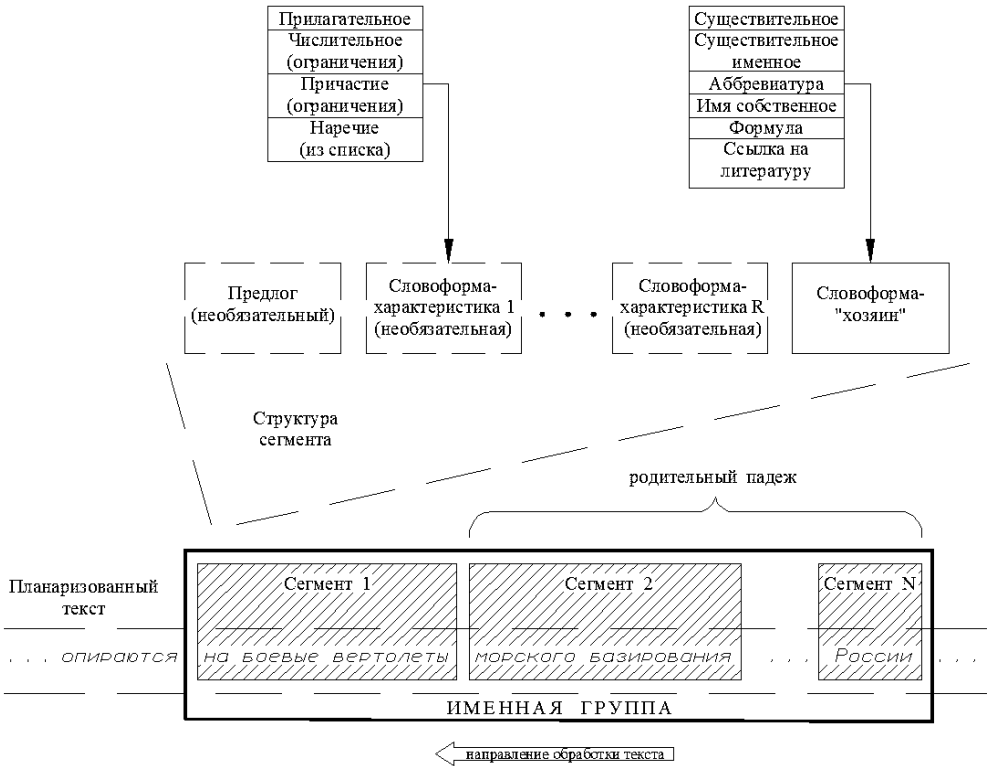

Рис. 2. Структура именной группы

В правилах указываются ограничения на тип главного слова именной группы (допускаются следующие типы: существительное, местоименное существительное, аббревиатура, имя собственное, формула или ссылка на цитату). Среди зависимых слов могут быть прилагательные, местоименные прилагательные, числительные, частицы, наречия и предлоги. Метод извлечения именных групп предназначен для поиска именных групп внутри предложения. Каждая именная группа может состоять из одного или нескольких сегментов, то есть групп слов с определенными характеристиками. Внутри сегмента все слова постоянны относительно своих грамматических характеристик. Если для именной группы построено несколько сегментов, то самый левый сегмент считается главным и может обладать произвольными грамматическими характеристиками (падежом и числом). Для остальных сегментов допустимой является только форма родительного падежа зависимого слова. Сбор сегментов именной группы производится справа налево. Во время аннотирования именной группы главное слово выделяется и нормализуется. Нормализованная форма именной группы помечается специальным атрибутом аннотирования «Form». Во время этой фазы могут возникать ошибки, связанные с омонимичностью определенных форм слов. Результаты аннотирования именных групп используются для онтологического аннотирования (при связывании с онтологией).

При построении именных групп математические выражения считаются корректными составляющими и могут быть включены в состав именной группы. Математическое выражение может выступать также в префиксной позиции, образуя дефисную конструкцию типа p-группа. С учетом формата разметки ArXMLiv, а также наличия специальной математической аннотации для математических выражений подобные компоненты разбираются сравнительно легко.

2.3. Семантическое аннотирование. Данная подсистема обеспечивает функциональность аннотирования документов в терминах заданного набора онтологий предметных областей.

Онтология Mocassin. Онтология[11] проекта Mocassin[12] предназначена для описания семантики структурных элементов, типичных для научных статей по математике. Каждый структурный элемент представляет собой фрагмент текста и характеризуется своим расположением в тексте, текстовым и формульным содержанием, а также уникальной функциональной нагрузкой. В частности, определяются такие стандартные части математических статей, как теоремы, леммы, доказательства, определения, следствия и т. д. Кроме того, онтология выделяет два вида бинарных отношений между структурными элементами – навигационный и ограниченный. Отношения первого вида, такие как *refersTo* ("*ссылаетсяНа*") и *dependsOn* ("*зависитОт*"), возникают при наличии в тексте ссылочных предложений на значительные части анализируемой или других статей (например, «используем *Лемму* 1 при доказательстве следующего факта...»). Отношения *hasPart* (часть – целое) и *followedBy* (отношение предшествования в тексте) также относятся к первому виду. Примером отношений второго типа является отношение *proves* ("*доказывает*"), которое возникает между доказательством (единственно возможным типом субъекта для данного отношения) и утверждением, которое требуется установить. Онтология использует элементы SALT Document Ontology[13] – онтологии риторических структур (структур изложения) научных публикаций. В частности, эта онтология определяет такие классы, как *Section* (*Глава*), *Figure* (*Рисунок*) и *Table* (*Таблица*).

Для создания связи между структурными элементами и другими объектами, содержащимися внутри структурных элементов или вне их, например, именованными математическими сущностями, извлеченными из текста, было введено специальное свойство *mentions* ("*упоминает*") вида:

$$mentions(x, y) \rightarrow (DocumentSegment(x) \vee Table(x) \vee Figure(x) \vee Section(x)) \wedge Thing(y)$$

Класс *DocumentSegment* (*Сегмент Документа*) является корнем иерархии онтологии Mocassin.

В онтологии определены также классы, представляющие различные виды математических выражений – *Mathematical Expression* (*Математическое Выражение*), *Variable* (*Переменная*) и *Formula* (*Формула*). Отношение *hasLatexSource* используется для представления связи между формулой (математическим выражением) и ее представлением на языке LaTeX. Однако для связывания формул с именованными математическими сущностями в онтологии существует отношение *hasNotation* вида: $hasNotation(x, y) \rightarrow Thing(x) \wedge MathematicalExpression(y)$.

Кроме того, онтология содержит аксиомы мощности, например одна из них декларирует, что каждое доказательство соотносится с не более чем одним утверждением, а также дополнительные логические правила, например, $dependsOn(x, y) \wedge hasPart(z, y) \rightarrow dependsOn(x, z)$.

Онтология разработана на языках OWL2/RDFS[14], которые обеспечивают богатые выразительные возможности, а также теоретические и практические средства вывода, например, с использованием современных машин вывода, таких как Pellet[15] и FaCT++[16].

[11] The ontology is accessible via URL: http://cll.niimm.ksu.ru/ontologies/mocassin (login/password: demo/demokpfu)
[12] http://code.google.com/p/mocassin/
[13] http://salt.semanticauthoring.org/ontologies/sdo
[14] http://www.w3.org/TR/owl2-rdf-based-semantics/
[15] http://clarkparsia.com/pellet/
[16] http://owl.man.ac.uk/factplusplus/

Извлечение логической структуры. Данный модуль получает результат обработки текста и выделяет структурные элементы в терминах онтологий Mocassin и SALT SDO по методу, предложенному в [11]. Этот метод предназначен для решения двух задач: (i) распознавание типов структурных элементов; (ii) распознавание семантических связей между ними. Результатом работы модуля является семантический граф, который содержит, с одной стороны, структурные элементы в качестве вершин, каждая из которых либо сопоставлена с определенным классом онтологии, либо отмечена как «нераспознанная», а с другой – экземпляры онтологических отношений в качестве ребер. Помимо функциональных свойств, каждая вершина аннотирована соответствующими свойствами типа данных, такими как заголовок, состав текста, номера страниц в скомпилированном PDF-документе.

OntoMath$^{\mathbf{Pro}}$. OntoMath$^{\mathrm{Pro}}$ – прикладная онтология для автоматической обработки математических статей на русском языке. Эта онтология определяет концепты, широко используемые в математике разного уровня, а также развивающийся лексикон профессиональной математики[17]. OntoMath$^{\mathrm{Pro}}$ охватывает широкий спектр областей математики, таких как теория чисел, теория множеств, алгебра, математический анализ, геометрия, математическая логика, дискретная математика, комбинаторика, дифференциальные уравнения, численный анализ, теория вероятности и математическая статистика. В метаданных каждого класса приводится его определение и наиболее употребляемые наименования (в основном на русском языке, но для некоторых классов даны и английские названия), включая синонимы.

В процессе разработки использовались следующие терминологические источники: классические книги, интернет-ресурсы, такие как Wikipedia и Cambridge Mathematical Thesaurus, научные статьи журнала «Известия вузов. Математика», а также личный опыт профессиональных математиков Казанского федерального университета.

Концепты онтологии организованы отношениями ISA в две таксономические структуры: иерархию областей математики и иерархию объектов математического знания. Первая иерархия является, в известном смысле, типовой и тесно связана с соответствующей частью классификатора УДК. Верхний уровень другой таксономии содержит концепты трех типов: (i) базовые метаматематические концепты, такие как, например, Множество, Оператор, Отображение, Тензор; (ii) корневые элементы концептов, относящихся к определенным областям математики, например, «Элементы теории вероятности» или «Элементы численного анализа»; (iii) классические научные концепты: Задача, Метод, Утверждение, Формула и т. п. В иерархии допускается множественное наследование (один и тот же класс может быть подклассом нескольких классов), например, класс «Формула Грина» является подклассом как класса «Формулы математического анализа», так и класса «Формулы численного анализа».

OntoMath$^{\mathrm{Pro}}$ определяет три типа свойств объектов:

- прямое отношение между объектом математического знания и областью математики *belongsTo* (“*принадлежитК*”);
- прямое отношение логической зависимости между объектами математического знания *isDefinedBy* (“*определяетсяПосредством*”);
- симметричное ассоциативное отношение («слабой зависимостью») между объектами математического знания *seeAlso* (*смТакже*).

Онтология OntoMath$^{\mathrm{Pro}}$ разработана на языках OWL-DL/RDFS и содержит 3450 классов, 6 типов свойств объектов, 3630 экземпляров свойства IS-A и 1140 экземпляров остальных свойств.

[17]Например, термин «задача Бицадзе-Самарского» - один из примеров сущностей, для которых невозможно найти в ресурсах типа Википедия соответствующей страницы

Извлечение именованных математических сущностей. Данный модуль рассматривает именные группы, полученные от подсистемы обработки текста в качестве кандидатов в экземпляры классов OntoMath$^{\text{Pro}}$, которые в дальнейшем будем называть именованными математическими сущностями (ИМС). Извлечение ИМС является нечетким и основывается на сравнении набора слов в именной группе и набора слов в названии класса онтологии. В качестве основы для меры уверенности используется хорошо известная мера Жаккара. Как следствие, метод предусматривает выбор доверительного порога для фильтрации неверных совпадений. Конкретно для заданной именной группы и класса онтологии мера уверенности C определяется по следующим правилам:

• C находится в интервале от 0 (минимальная уверенность) до 1 (максимальная уверенность);

• если название класса не содержит главное слово именной группы (NP.Head), то $C = 0$;

• если длина (число слов) названия класса превышает длину именной группы, то $C = 0$;

• иначе C равно коэффициенту близости Жаккара для наборов слов.

Например, мера уверенности между строками «Пространство типа Соболева» (строка из текста) и «Пространство Соболева» (название класса онтологии) равна 2/3. С другой стороны, мера уверенности между строками «число» (строка из текста) и «число Ферма» (название класса онтологии) равна 0 из-за разных длин, а мера уверенности между строками «интеграл от функции комплексной переменной» (строка из текста) и «функция комплексной переменной» (название класса онтологии) также равна 0 из-за несовпадения главных слов в этих фразах.

Модуль добавляет семантические аннотации к именным группам в виде URI установленного класса онтологии.

Сопоставление ИМС с формулами. Данный модуль решает следующие задачи анализа внутри документа:

• разбор математических выражений, включающий определение переменных и их поиск в математических формулах;

• сопоставление математических выражений с именными группами.

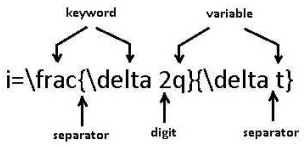

Рис. 3. Структура математического выражения

Модуль опирается на математические аннотации, токены, предложения и именные группы, полученные на предыдущих этапах. В частности, исходные строки LaTeXв математических аннотациях анализируются для получения более точного представления формул (рис. 3). В качестве инструмента анализа формул используются регулярные выражения. Сначала формула очищается от специальных элементов разметки и избыточных пробелов. Формула разделяется на отдельные элементы, разделителями являются скобки, квадратные скобки, операторы, знаки пунктуации и пробелы. Данные элементы относятся к специальным группам – ключевым словам (стандартные команды LaTeX), индексам, числам и т. п. Для каждого неклассифицированного элемента дополнительно выполняются проверки некоторых условий, например, являются ли начальные символы числовыми подстроками

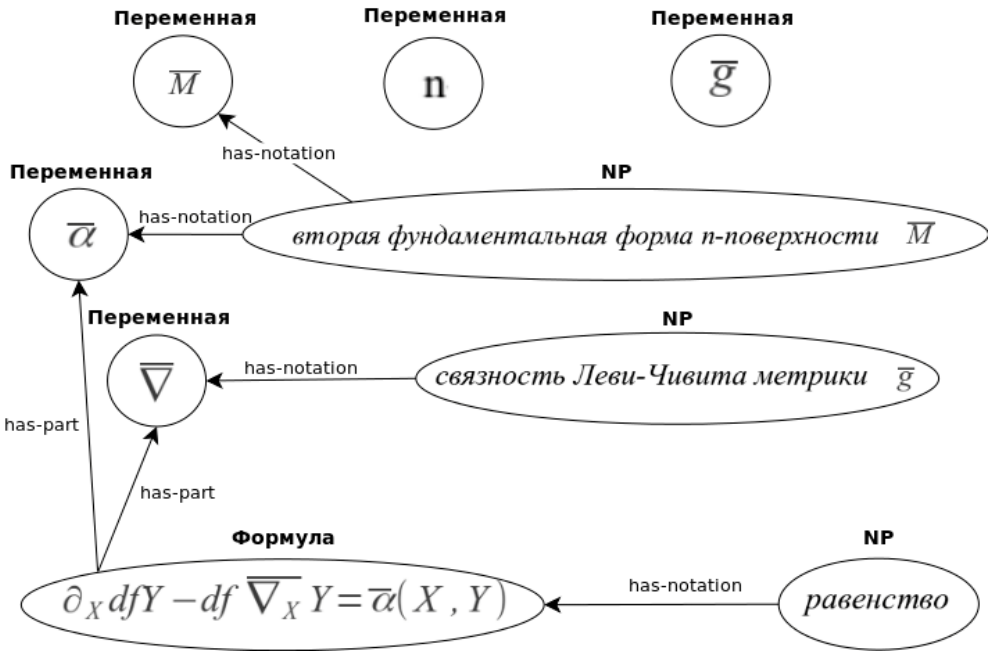

Рис. 4. Семантический граф примера 1

или принадлежит ли элемент к множеству греческих символов. В результате математические выражения разделяются на три группы: переменные, формулы и вспомогательные фрагменты.

Все переменные и формулы сохраняются в специальном индексе, который содержит информацию о появлении переменных в формулах. Приведем пример, иллюстрирующий семантику подобных отношений.

Пример 1. Рассмотрим фрагмент математического текста:

Пусть $\overline{\alpha}$ – вторая фундаментальная форма n-поверхности $\overline{M}$, $\overline{\nabla}$ – связность Леви–Чивита метрики $\overline{g}$, тогда выполняется равенство:

$$\partial_X dfY - df\overline{\nabla}_X Y = \overline{\alpha}(X, Y).$$

Фрагмент текста содержит переменные $\overline{\alpha}$, n, $\overline{M}$ и $\overline{g}$; формула использует переменные $\overline{\alpha}$ и $\overline{\nabla}$. Переменные X и Y не определены в документе и поэтому отсутствуют в индексе. Экземпляры отношения *has-part* проиллюстрированы на рис. 4.

Следующий этап заключается в связывании именных групп с извлеченными формулами и переменными (в дальнейшем все описания, касающиеся связывания формул, применимы в полном объеме и к связыванию переменных). Можно выделить два возможных случая взаимного расположения формулы и именной группы: во-первых, именная группа может содержать формулу; во-вторых, элементы (формула и именная группа) могут следовать друг за другом.

В первом случае именная группа является единственным кандидатом для связывания. В простейшем случае именная группа состоит из единственного главного слова. В примере выделена именная группа «равенство $», где $ – формула в составе именной группы. В дальнейшем символ $ будет заменять конкретные математические выражения в текстовых примерах. В этом случае происходит автоматическое связывание формулы с данной именной группой (см. рис. 4). В более

сложном случае именная группа содержит более чем одно слово. В этом случае рассматривается расстояние между формулой и главным словом. Если это расстояние составляет более трех слов (эмпирическая оценка), то формула считается дополнением и не связывается. В примере переменная $\overline{g}$ не связывается с именной группой *«связность Леви–Чивита метрики $»* по этому критерию. Префиксы математических выражений игнорируются – в примере префикс n- не связывается (n-поверхность), однако переменная $\overline{M}$ связывается с именной группой *«вторая фундаментальная форма $n-$поверхности $\overline{M}$»*.

Основной идеей анализа во втором случае является концепция максимально возможной дистанции (MFD) в терминах позиций символов между математическими аннотациями и аннотациями именных групп в тексте. Для заданной пары MFD предполагается меньше длины предложения, которое содержит обе аннотации. Оптимальное значение MFD, определенное эмпирически, равно 25 символам. Наконец, метод позволяет определять ближайшую аннотацию именной группы для заданной формулы. При этом некоторые случаи анализируются специальным образом, например, часто встречающиеся паттерны, такие как «Формула – Именная группа» (со знаком тире между элементами).

Результатом работы модуля является семантический граф с дополнительными аннотациями.

2.4. Извлечение метаданных статьи. Данный модуль извлекает метаданные статьи и библиографические ссылки в терминах онтологии AKT (AKT Portal Ontology) [12]. В модуле извлечения выполняются следующие процедуры:

- просмотр коллекции документов и извлечение метаданных из заголовков – название, имена авторов и их места работы, название журнала, номер журнала и год публикации;
- создание идентификаторов для опубликованных статей;
- обработка библиографических описаний статей, используя построенные идентификаторы.

Затем данные подготавливаются к сериализации согласно схеме AKT:

- даты представляются как экземпляры класса akt:Calendar-Date;
- данные журнала группируются в экземпляр класса akt:Journal со свойствами akt:has-date и akt:has-issue-number;
- Места работы отображаются в экземпляры класса akt:Organization;
- данные автора представляются при помощи экземпляров класса akt:Affiliated-Person: свойство akt:full-name используется для имени автора, свойство akt:has-affiliation – для установления связи с соответствующей организацией, свойство akt:has-email-address – для хранения адреса электронной почты;
- статьи соответствуют экземплярам класса akt:Article-Reference со свойствами akt:has-title, akt:has-author, akt:included-in-publication;
- библиографические ссылки вводятся посредством свойства akt:cites-publication-reference.

Все URI экземпляров класса Article-Reference создаются совместимыми с крупной интернет-коллекцией MathNet.Ru[18]. В частности, это означает, что данные URI могут быть разыменованы в браузере посредством соответствующего ресурса.

2.5. Создание RDF. В данном модуле используются все данные, полученные на предыдущих шагах, для их преобразования в RDF-представление. Для этой

[18] http://mathnet.ru

задачи используется библиотека OpenRDF Sesame[19], написанная на Java, с помощью которой создаются RDF-триплеты. Построенные триплеты сохраняются в специализированном хранилище – экземпляре сервера Virtuoso Community Edition[20], который представляет собой высокопроизводительную СУБД с поддержкой RDF/SPARQL и возможностями логического вывода.

2.6. Связывание. Данная подсистема не связана с модулями обработки, описанными выше, и, в отличие от них, ее применение требует значительных усилий со стороны человека. Подсистема решает задачу связывания набора данных IVM с существующими наборами данных в облаке LOD. По-существу, при связывании необходимо решить две задачи: во-первых, выполнить связывание онтологии $\text{OntoMath}^{\text{Pro}}$ с ресурсами DBPedia, а во-вторых, осуществить связывание ресурсов из набора IVM с соответствующими ресурсами из наборов облака LOD, основанных на схеме AKT.

Для решения задач связывания использовалась система Silk[21].

Связывание классов онтологии $\text{OntoMath}^{\text{Pro}}$ с ресурсами DBPedia. При проведении связывания используются следующие свойства ресурсов:

- названия классов онтологии и ресурсов DBPedia (свойства rdfs:label);
- ссылки на Википедию. При разработке онтологии некоторые определения классов были импортированы из Википедии и поэтому содержат ссылки на соответствующие страницы. Эти ссылки сравниваются со значениями свойств foaf:primaryTopic и rdfs:labels ресурса DBPedia, используя интервики-ссылки (ссылки между языковыми разделами) для русскоязычных статей Википедии.

Для связывания применяются лишь те ресурсы DBPedia, которые принадлежат к категории «Математика» или ее подкатегориям (например, Алгебра, Геометрия, Математическая Логика), вплоть до четвертого уровня вложенности свойства skos:broader.

Это ограничение вызвано выявленными недостатками системы Silk и ресурса DBPedia, касающимися создания и представления транзитивных свойств[22].

После проведения связывания создаются триплеты, соединяющие классы онтологии $\text{OntoMath}^{\text{Pro}}$ с ресурсами DBPedia при помощи свойства skos:closeMatch.

Поиск дубликатов в наборах LOD, основанных на схеме AKT. Посредством программной утилиты RKBExplorer[23] были исследованы наборы данных, основанные на схеме AKT. Было установлено, что наиболее подходящим набором данных для поиска дубликатов является набор CORDIS[24]. Связывание осуществлялось на основе информации об организациях, в частности, использовались такие свойства, как akt:name и akt:has-pretty-name.

Результаты связывания подробно обсуждаются в разд. 4.

3. Реализация модулей прототипа

Подсистема обработки текста базируется на компонентах системы «OntoIntegrator» [13] – программного комплекса лингвистического анализа русскоязычных текстов, поддерживающего XML в качестве формата ввода/вывода.

[19] http://www.openrdf.org/

[20] http://sourceforge.net/projects/virtuoso/

[21] http://www4.wiwiss.fu-berlin.de/bizer/silk/

[22] Заметим, что во время экспериментов с большим числом уровней иерархии свойства skos:broader были получены некорректные результаты. Например, была построена транзитивная цепочка, связывающая категории «Топология» и «Алиса в Стране чудес» и др.

[23] http://www.rkbexplorer.com/explorer

[24] http://thedatahub.org/dataset/rkb-explorer-cordis

Модуль извлечения ИМС реализован в виде JS-скрипта[25], принимающего на вход OWL-файл онтологии OntoMath$^{\text{Pro}}$ и XML-файлы математических документов для обработки. Скрипт, основываясь на результатах аннотирования документа системой «OntoIntegrator», дополняет XML-файлы новыми атрибутами.

Модуль извлечения логической структуры является частью проекта Mocassin, открытого движка семантического поиска, написанного на языке Java. Модуль обрабатывает XML-документы, основываясь на системе аннотирования, принятой в архитектуре платформы GATE[26] и ряда специальных анализаторов.

Модуль связывания ИМС с формулами реализован в виде расширения GATE[27].

Каждый программный модуль в составе рассматриваемой программной платформы использует аннотации, полученные на предыдущих стадиях обработки документа.

Извлечение метаданных статей выполняется специальными Shell-скриптами[28], а связывание реализуется специальным скриптом, запускающим систему Silk и выделяющим список категорий ресурса DBPedia, относящихся к математике[29].

4. Эксперименты и оценка результатов

В данном разделе представлены результаты проведенных экспериментов, выполнено оценивание полученных данных и изложен анализ возможных недостатков предложенных методов.

Обработка текста. Из коллекции были случайным образом отобраны 10 документов, для которых все аннотации (общее число построенных аннотаций около 3000) проверялись вручную. Авторы реальных математических статей при их оформлении часто допускают ряд некорректностей, которые влияют на точность результата сегментации текстов. Для корректных текстов точность сегментации предложений составляет 99%; в случае тех или иных некорректностей оформления точность результата может снижаться. Некорректные тексты содержат, например, точку в конце формулы внутри тега <Math>, или не содержат финальной точки после формулы. В этом случае, метод сегментации пытается привязать границу предложения к другим элементам, например, к тегу абзаца <para>. При экспериментальной проверке метода в тестируемой математической коллекции обнаружены некорректные (по вине автора) тексты, для которых точность сегментации предложений составила не менее 95%.

Метод выделения именных групп дает точность не ниже 88%. Основными типами ошибок являются пропущенные устойчивые предложные фразы (5%), пропущенные правые части определений (2%), неполные структуры именных групп (2%) и т. д. Результаты данного метода могут быть улучшены за счет более глубокого синтаксического анализа и использования большего числа шаблонов (устойчивых словосочетаний, используемых в математических текстах).

Извлечение ИМС. При индексировании всей коллекции математических текстов на этапе обработки текстов было построено 330462 аннотаций именных групп. Модуль извлечения ИМС связал 138032 (41.7%) именных групп с классами онтологии с ненулевой мерой уверенности. После удаления дубликатов было получено 16300 уникальных кандидатов в ИМС, которые были сгруппированы по значениям меры уверенности и переданы эксперту-математику для экспертизы. В табл. 1

[25] http://bit.ly/cll-mne-extraction
[26] http://gate.ac.uk/
[27] http://bit.ly/cll-gate-morph-formula
[28] http://bit.ly/cll-akt-metadata-extraction
[29] http://bit.ly/cll-interlinking

Табл. 1

Оценка извлечения ИМС

Порог меры уверенности	Число кандидатов	Число корректных кандидатов	% от общего	Точность
0.27	16300	12255	1.000	0.752
0.29	16296	12254	1.000	0.752
0.33	**15964**	**12117**	**0.989**	**0.759**
0.36	8312	6073	0.496	0.731
0.38	8311	6072	0.495	0.731
0.40	8281	6054	0.494	0.731
0.43	7316	5388	0.440	0.736
0.44	7277	5374	0.439	0.738
0.45	7275	5374	0.439	0.739
0.50	7274	5373	0.438	0.739
0.57	**2470**	**2426**	**0.198**	**0.982**
0.60	2466	2422	0.198	0.982
0.67	2384	2357	0.192	0.989
0.71	1266	1266	0.103	1.000
0.80	1265	1265	0.103	1.000
1.00	1254	1254	0.102	1.000

Табл. 2

Оценка метода связывания ИМС с формулами

	Формула	Именные группы	Определения переменных	Число переменных в формулах	Истинно-положительные	Истинно-отрицательные	Ошибки	Истинные результаты
Всего	1562	1745	177	581	672	450	440	1122
Min-Max по документу	144–165	109–285	8-29	14–101	36–85	229–63	29–65	91–136
% общего числа формул	100	–	11	–	43	29	29	72

показано распределение оценок полноты/точности результатов в зависимости от меры уверенности.

В результате для построения набора данных RDF была выбрана следующая стратегия:

• для кандидатов с мерой уверенности более 0.57 создаются экземпляры «жестких» отношений (rdf:type), в которых каждая именная группа считается экземпляром классов связанной онтологии;

• для кандидатов с мерой уверенности между 0.33 и 0.57 создаются экземпляры «мягких» отношений (skos:closeMatch).

Связь ИМС с формулами. Для экспериментов вручную были проанализированы 10 документов из различных областей математики. При анализе учитывалось количество фрагментов формул, содержащихся в каждом документе. Общая оценка результатов экспериментов представлена в табл. 2.

Всего в документах было выделено 1562 математических выражения, из которых 117 (11%) представляли собой выражения для переменных. Общее число описаний переменных – 581. Посредством метода корректно связано (истинно-

положительно) 43% формул и не связано (истинно-отрицательно) 29%, что обеспечивает оценку точности 72%. Следует отметить, что данная оценка точности получается с учетом погрешностей обработки текста и некоторых орфографических неточностей, таких как замена знака тире дефисами при наборе формул. Среди предполагаемых улучшений метода можно указать более точную настройку максимально возможной дистанции (MFD), специальную обработку групп математических выражений и улучшение фильтрации математических выражений.

Связывание классов онтологии OntoMath$^{\text{Pro}}$ с ресурсами DBPedia. В результате связывания установлены 842 связи с участием 828 классов онтологии OntoMath$^{\text{Pro}}$ (некоторые классы связывались с несколькими ресурсами DBPedia). Таким образом, оценка покрытия онтологии OntoMath$^{\text{Pro}}$ составляет 24%, а ручная оценка точности – 95%. Ошибки возникали в следующих случаях:

• неточные ссылки в комментариях онтологии, например, E203 Компактный Слой ≠ dbpedia:Novikov's_compact_leaf_theorem;

• некорректные интервики-ссылки в Википедии, например, E3263 Сумма последовательности ≠ dbpedia:Convergence_tests;

• существование в DBPedia омонимичных ресурсов, например E1408 Отображение спаривания ≠ dbpedia:Mating (Сексуальное спаривание), которое попало в категорию Статистика через следующую цепочку подкатегорий: Фертильность → Демография → Разделы и приложения статистики → Статистика

Поиск дубликатов в наборах данных, базирующихся на схеме AKT. В результате экспериментов по связыванию организаций из набора IVM с соответствующими организациями из набора CORDIS удалось найти 91 корректных и 13 некорректных связей. Связывание с организациями из набора DBLP не дало результатов, возможно, по причине ограничений его SPARQL-точки доступа.

5. Набор данных IVM: сценарии использования

RDF-набор данных IVM, полученный в результате обработки коллекции математических статей журнала «Известия вузов. Математика», содержит 854284 триплетов, включая описания 4190 теорем, 3035 доказательств, 2356 лемм, 1015 определений и других индексированных математических сущностей. Ниже приведены ряд примеров использования набора IVM в практических приложениях поиска. Примеры представляют собой формулировки SPARQL-запросов, иллюстрирующих возможные практические приложения.

Пример 2. Найти все статьи, содержащие теоремы о конечных группах.

```
PREFIX moc: <http://cll.niimm.ksu.ru/ontologies/mocassin#>
PREFIX math: <http://cll.niimm.ksu.ru/ontologies/mathematics#>
SELECT ?article WHERE {
?article moc:hasSegment ?theorem .
?theorem moc:mentions ?o .
?theorem a moc:Theorem .
?o a math:E2183
}
```

В заголовке запроса определены префиксы онтологии, в теле запроса используется класс «Теорема» онтологии Mocassin, а также класс «Конечная группа», свойства hasSegment и mentions онтологии OntoMath$^{\text{Pro}}$.

Пример 3. Определить области математики, к которым относится данная статья.

```
define input:inference "http://cll.niimm.ksu.ru/ontologies/mathematics/
rules"
PREFIX moc: <http://cll.niimm.ksu.ru/ontologies/mocassin#>
PREFIX math: <http://cll.niimm.ksu.ru/ontologies/mathematics#>
SELECT ?field WHERE {
<http://mathnet.ru/ivm327> moc:hasSegment ?element .
?element moc:mentions ?entity . ?entity a ?entityClass .
?entityClass owl:equivalentClass ?r . ?r owl:onProperty math:P3 .
?r owl:allValuesFrom ?field .
} GROUP BY ?field
```

В заголовке указаны префиксы используемых онтологий и раздел логического вывода онтологии OntoMath$^{\text{Pro}}$. Для поиска задана статья из ресурса Mathnet с URI http://mathnet.ru/ivm327. Результатами данного запроса являются классы, представляющие области математики, такие, как дискретная математика, комбинаторика, математический анализ и теория вероятностей.

Заключение

В работе представлен прототип программной платформы для извлечения структурированных стандартизованных представлений научных статей по математике. Система-прототип может быть использована для публикации метаданных и содержимого математических статей в формате, совместимом с данными LOD. Разработанный прототип был апробирован на коллекции математических статей (общий объем коллекции более 1300 статей) для демонстрации возможностей предложенного подхода. Представленные результаты и их оценки позволяют делать вывод об эффективности принятых решений. Особое внимание также уделено рассмотрению примеров использования построенного набора данных в практических приложениях математического поиска.

Авторы выражают особую благодарность математикам Казанского федерального университета – разработчикам онтологии OntoMath$^{\text{Pro}}$: В.Д. Соловьеву, А.В. Каюмовой, И.Р. Каюмову, П.Н. Иваньшину, Е.К. Липачеву, М.С. Матвейчуку, Е.А. Уткиной.

Работа выполнена при финансовой поддержке Минобрнауки России (государственный контракт от 20.10.2011 г. № 07.524.11.4005 в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 годы») и РФФИ (проект № 11-07-00507-а).

Summary

O.A. Nevzorova, N.G. Zhiltsov, D.A. Zaikin, O.N. Zhibrik, A.V. Kirillovich, V.N. Nevzorov, E.V. Biryaltsev. A Software Platform Prototype for Publishing Semantic Data from Scientific Collections in Mathematics to the LOD Cloud.

This paper deals with a software platform prototype for extraction of Linked Open Data (LOD) from a given collection of mathematical scholarly papers. The problem of obtaining the semantic representation of a collection in the chosen subject area is of topical interest since the LOD cloud currently lacks up-to-date data on professional mathematics. We believe that the main reason for that is the absence of appropriate tools that could analyze the underlying semantics in mathematical papers and effectively build their consolidated representation. In this article, we describe a complex approach to the analysis of these documents for representing their content and metadata in RDF format. We also consider methods and

techniques based on special ontologies for extracting semantic data from mathematical papers and describe experiments on integration of the constructed RDF-set into the existing datasets on the Internet.

Key words: indexing, linked data, ontology engineering.

Литература

1. *Muhleisen H., Bizer C.* Web Data Commons – Extracting Structured Data from Two Large Web Corpora // Proc. WWW2012 Workshop on Linked Data on the Web. – 2012. – URL: http://www.wiwiss.fu-berlin.de/en/institute/pwo/bizer/research/publications/MuehleisenBizerWebDataCommonsLdow2012.pdf, свободный.
2. *Kamareddine F., Wells J.B.* Computerizing mathematical text with MathLang // Electr. Notes Theor. Comp. Sci. – 2008. – V. 205. – P. 5–30.
3. *Kohlhase M.* OMDoc – An Open Markup Format for Mathematical Documents [Version 1.2]. – Berlin: Springer-Verlag, 2006. – 428 p.
4. *David C., Kohlhase M., Lange C., Rabe F., Zhiltsov N., Zholudev V.* Publishing Math Lecture Notes as Linked Data // Proc. 7th Extended Semantic Web Conf. (LNCS No 6089). – Berlin: Springer-Verlag, 2010. – P. 370–374.
5. *Baker C., Kanagasabai R., Ang W., Veeramani A., Low H., Wenk M.* Towards ontology-driven navigation of the lipid bibliosphere // BMC Bioinformatics. – 2008. – V. 9, Suppl. 1. – URL: http://www.biomedcentral.com/1471-2105/9/S1/S5, свободный.
6. *Volker J., Haase P., Hitzler P.* Learning Expressive Ontologies // Bridging the Gap between Text and Knowledge. – Amsterdam: IOS Press, 2008. – P. 45–69.
7. *Volker J., Fernandez-Langa S., Sure Y.* Supporting the Construction of Spanish Legal Ontologies with Text2onto // Computable Models of the Law. – Berlin: Springer-Verlag, 2008. – P. 105–112.
8. *Liu W., Jin F., Zhang X.* Ontology-Based User Modeling for E-Commerce System // Proc. 3rd Int. Conf. on Pervasive Computing and Applications (ICPCA). – 2008. – V. 1. – P. 260–263.
9. *Бирюльцев Е.В., Елизаров А.М., Жильцов Н.Г., Иванов В.В., Невзорова О.А., Соловьев В.Д.* Модель семантического поиска в коллекциях математических документов на основе онтологий // Труды 12-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2010. – Казань: Казан. гос. ун-т, 2010. – С. 296–300.
10. *Stamerjohanns H., Kohlhase M., Ginev D., David C., Miller B.* Transforming Large Collections of Scientific Publications to XML // Math. Comp. Sci. – 2010. – V. 3. – P. 299–307.
11. *Solovyev V., Zhiltsov N.* Logical Structure Analysis of Scientific Publications in Mathematics // Proc. Int. Conf. on Web Intelligence, Mining and Semantics (WIMS'11). – N. Y.: ACM Press, 2011. – No 21. – P. 1–9.
12. *Schraefel M., Shadbolt N., Gibbins N.* CS AKTive Space: Representing Computer Science on the Semantic Web // Proc. WWW 2004. – N. Y.: ACM Press, 2004. – P. 384–392.
13. *Nevzorova O., Nevzorov V.* The Development Support System "OntoIntegrator" for Linguistic Applications // Information Science and Computing. – Rzeszow-Sofia: ITHEA, 2009. – V. 3, No 13. – P. 78–84.

Поступила в редакцию
09.07.12

Невзорова Ольга Авенировна – кандидат технических наук, доцент, заместитель директора НИИ «Прикладная семиотика» АН РТ, доцент кафедры информационных систем Казанского (Приволжского) федерального университета.

E-mail: *onevzoro@gmail.com*

Жильцов Никита Геннадьевич – младший научный сотрудник лаборатории интеллектуального поиска и семантических технологий ВШИТИС Казанского (Приволжского) федерального университета.

E-mail: *nikita.zhiltsov@gmail.com*

Заикин Данила Александрович – аспирант кафедры теоретической кибернетики Казанского (Приволжского) федерального университета.

E-mail: *ksugltronteal@gmail.com*

Жибрик Ольга Николаевна – научный сотрудник лаборатории технологии баз данных НИЦ «НИИММ им. Н.Г. Чеботарева» Казанского (Приволжского) федерального университета.

E-mail: *olgazhibrik@gmail.com*

Кириллович Александр Витальевич – аспирант кафедры теоретической кибернетики Казанского (Приволжского) федерального университета.

E-mail: *Alik.Kirillovich@gmail.com*

Невзоров Владимир Николаевич – кандидат технических наук, доцент кафедры информационных технологий проектирования электронно-вычислительных средств Казанского национального исследовательского технического университета имени А.Н. Туполева.

E-mail: *nevzorovvn@gmail.com*

Биряльцев Евгений Васильевич – кандидат технических наук, заведующий лабораторией технологии баз данных НИЦ «НИИММ им. Н.Г. Чеботарева» Казанского (Приволжского) федерального университета.

E-mail: *IgenBir@yandex.ru*