О.А. Невзорова, Е.В. Биряльцев, Н.Г. Жильцов

# Коллекции математических текстов: аннотирование и применение в поисковых задачах[1]

**Аннотация.** В статье рассматриваются модель семантического аннотирования математических текстов и модель семантического поиска в размеченной коллекции математических текстов. Обсуждаются результаты экспериментов по семантическому аннотированию коллекции научных публикаций по математике.

**Ключевые слова:** семантическое аннотирование, семантическая аннотация, математическая коллекция.

## Введение

Семантический веб активно развивает онтологический подход, при котором онтологии рассматриваются как ресурсы для аннотирования документов. Аннотированные документы широко используются компьютерными системами для извлечения знаний из текстов [1], а также позволяют обрабатывать сложные пользовательские запросы при информационном поиске.

Поиск в коллекциях математических документов [2] – актуальная и быстроразвивающаяся область исследований. Современные поисковые системы для научных коллекций условно можно разделить на три группы. К первой относятся системы поиска научных публикаций [3, 4] и поисковые интерфейсы крупнейших научных коллекций [5-7], которые предлагают сервис полнотекстового поиска по ключевым словам с учетом метаданных публикации (автор, название, журнал, краткое описание). Эти системы индексируют значительные объемы актуальных научных статей в области математики в формате PDF или LaTeX. Отличительная особенность систем второй группы состоит в том, что они используют семантику математической нотации и реализуют поиск по формулам и выражениям [8-10]. Данные системы работают со специальным семантическим представлением математических формул, выраженным на языках Content/Presentation MathML и OpenMath. В качестве результатов возвращаются ссылки на документы, содержащие релевантные формулы. Согласно [8] основные трудности при разрешении запросов данного типа: зависимость форм нотации от контекста; неоднозначность трактовки одних и тех же символов; идентичность формул с точностью до обозначений. Третья группа систем – приложения, в основном, системы автоматических доказательств, использующие строго формализованное представление математических документов на таких языках как Mizar и Coq [11, 12]. Они позволяют выполнять сложные семантические запросы, например, искать теоремы для применения в данном доказательстве.

Общее направление развития современных поисковых технологий ориентировано на широкое привлечение семантики, учитывающей особенности предметной области, для повышения качественных характеристик поиска. В настоящей статье рассматривается расширенная модель представления математического документа, основанная на специализированных моделях семантических аннотаций, которая, по существу,

[1] Работа выполнена при поддержке РФФИ (проекты № 09-07-12059-офи_м, № 11-07-00507-а), Минобрнауки России по государственному контракту 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».

позволяет дать формальное описание семантической модели представления документа. Предложенный подход подразумевает интегрирование различных семантических аннотаций на уровне документа и может быть использован для построения эффективных технологий поиска и интеллектуальной обработки исходных текстов.

## 1. Модель математического документа

Одной из самых распространенных моделей представления документов в задачах информационного поиска является векторная модель документа. Размерность этого вектора, как и размерность пространства, равна числу различных слов во всей коллекции, и является одинаковой для всех документов. Поисковый запрос представляется как вектор того же пространства, и соответствие документов запросу вычисляется на основе сравнения векторов по какой-либо, например, косинусной мере.

Однако применение векторной модели для математических документов имеет ряд очевидных недостатков. Главный из них — игнорирование важных характеристик математических текстов: наличие логической структуры и элементов математической нотации. Другой существенной характеристикой математических текстов является строгость и лаконичность математического языка, использование стандартных оборотов, характерных для различных структурных элементов математической статьи (введения, определений, теорем и др.). Это позволяет рассчитывать на успешное применение специальных лингвистических техник (специальных грамматик или лексико-синтаксических шаблонов) для анализа научных текстов.

Расширенная семантическая модель математического документа может быть представлена как агрегирование различных типов семантических аннотаций, наиболее существенными из которых являются: семантическая аннотация логической структуры документа; терминологическая аннотация; расширенная формульная аннотация.

### 1.1. Семантическая аннотация логической структуры документа

Моделирование логической структуры математических документов исследовалось в целях различных приложений. В работе [13] предложен формат ABCDE для публикации трудов научных конференций и семинаров, который характеризуется спецификацией следующих разделов документа: Annotation (аннотация), Background (теоретические основы), Contribution (вклад), Discussion (обсуждение) и Entities (список литературы). Формат другого проекта - SALT [14] - может рассматриваться как расширение формата ABCDE, который содержит не только онтологические модели для представления структуры научной публикации, но и инструменты для редактирования семантических аннотаций для статей в формате LaTeX. Онтология SALT Document Ontology определяет семантику отдельных сегментов документа, характерных для любой научной статьи (разделов, абзацев, рисунков, таблиц и т.д.), а также содержит отношения между сегментами, в основном, типа “часть-целое”. Кроме того, онтология SALT Rhetorical Ontology описывает семантику элементов риторической структуры и аргументации, отраженной в тексте типовой публикации.

Подход, предложенный в [15], использует выделение элементов логической структуры и ссылок между ними для улучшения навигации при чтении математических документов (например, для перехода к формулировке теоремы по ссылке в доказательстве).

Онтология HELM [16, 17] была одной из первых попыток представления структуры математического документа в контексте задачи семантического поиска. Однако данный проект не получил существенного развития, в отличие от формата OMDoc [18] и ассоциированной с ним онтологии OMDoc [19], выраженной на языке OWL-DL. Логической структуре документа в формате OMDoc отвечает уровень утверждений. Онтология OMDoc формально описывает отношения между элементами, например, "доказательство доказывает теорему`` или "пример относится к определению``. На Рис.1 представлен фрагмент онтологии OMDoc, представляющий структурные элементы и отношения между ними.

На основе тестирования существующих моделей представления математических документов в реальных научных коллекциях (прежде всего онтологий OMDoc и SALT) нами предложена онтология структуры научных публикаций по математике Mocassin, которая содержит типовые концепты и отношения и эффективно извлекаемые из текстов автомати-

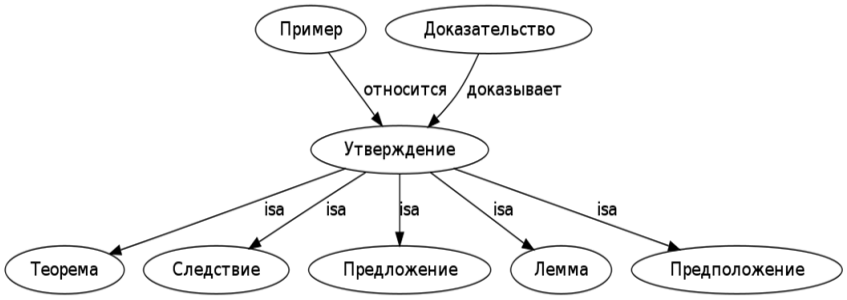


Рис. 1. Фрагмент онтологии OMDoc

ческими методами [20]. Онтология включает следующие концепты: DocumentSegment (Сегмент документа), Axiom (Аксиома), Claim (Утверждение), Conjecture (Гипотеза), Corollary (Следствие), Definition (Определение), Equation (Формула), Example (Пример) и др.

Математический документ в разработанной модели рассматривается как набор связанных сегментов, где сегмент является частью документа, имеет начальную и конечную позицию в тексте, а также характеризуется конкретной функциональной ролью. Онтология структуры научных публикаций по математике описывает семантику сегментов и возможных отношений между ними (Табл. 1). Важной характеристикой разработанной модели является ее интеграция с существующими моделями представления научных публикаций (в частности, с SALT).

Концепты с префиксом sdo - элементы онтологии SALT Document Ontology. Для интеграции семантических моделей вводится свойство hasSegment, связывающее понятия *публикации* и *сегмента документа*. Полное описание онтологии структуры научных публикаций по математике на языке OWL доступно по адресу: http://cll.niimm.ksu.ru/ontologies/mocassin.

Степень покрытия концептов онтологии в реальных математических текстах изучалась на двух тестовых коллекциях статей по математике в формате LaTeX. Первая коллекция представляла собой выборку из 1031 документа англоязычной коллекции arXiv.org. Вторая коллекция содержала 1355 статей журнала “Известия вузов. Математика” на русском языке за период с 1997 по 2009 гг.

Тестирование было выполнено на основе предобработки, которая включала следующие фазы: фильтрация элементов документа, не имеющих отношение к логической структуре (например, LaTeX -теги align, itemize, center); распознавание семантики тэгов на основе алгоритма близости строк q-gram и учета синонимии. В результате было получено 103680 и 40805 структурных элементов из первой и второй коллекции соответственно. Большая часть этих элементов (69% и 52%) - экземпляры концепта Equation (формула). Таким образом, доли концептов онтологии составили 91.6% и 91.9% (или 99.6% и 99.5% с учетом концепта Section из онтологии SALT) для обеих коллекций соответственно. Таким образом, экспериментальные данные подтверждают обоснованность выбора

Табл. 1. Отношения онтологии Mocassin

| Отношение | Наименование | Домен | Диапазон |
|---|---|---|---|
| dependsOn | зависит от | Claim, Conjecture, Corollary, Definition, Example, Lemma, Proof, Proposition, Theorem | Axiom, Claim, Conjecture, Corollary, Definition, Equation, Lemma, Proposition, Theorem |
| exemplifies | является примером для | Example | Axiom, Claim, Conjecture, Corollary, Definition, Equation, Lemma, Proposition, Remark, Theorem |
| hasConsequence | имеет следствие | Axiom, Claim, Conjecture, Corollary, Lemma, Proposition, Theorem | Corollary |
| hasSegment | содержит как сегмент | sdo:Publication | DocumentSegment |
| proves | доказывает | Proof | Claim, Corollary, Lemma, Proposition, Theorem |
| refersTo | ссылается на | sdo:Section, Axiom, Claim, Conjecture, Corollary, Definition, Example, Lemma, Proof, Proposition, Remark, Theorem | sdo:Section, Axiom, Claim, Conjecture, Corollary, Definition, Equation, Example, Lemma, Proof, Proposition, Remark, Theorem |

множества концептов данной онтологии для анализа логической структуры математических публикаций.

## 1.2. Терминологическая аннотация математического документа

Основными компонентами данной аннотации являются описания математических терминов. Для представления терминологии предлагается использовать подход на основе контролируемых словарей, которые специфицируют математические понятия и отношения между ними. Примерами таких ресурсов являются набор данных DBPedia [21] и математический тезаурус Кембриджского Университета [22].

Терминологическая аннотация включает не только границы терминологического словосочетания, но и структурные параметры (тип синтаксической связи и синтаксические роли компонентов словосочетания). Модели терминологических словосочетаний выделяются на основе моделей синтаксических отношений в русском языке. Обобщенной синтаксической моделью для терминологического словосочетания является именная группа (ИГ). В составе ИГ выделяется синтаксический хозяин (главное слово - имя существительное) и различные модификаторы (зависимые слова). Структура ИГ определяется набором синтаксических отношений между главным и зависимыми словами. Синтаксические отношения между компонентами словосочетаний (главным и зависимыми словами) строятся на основе взаимодействия лексических значений этих компонент и их грамматических форм. Все многообразие этих отношений в русском языке обобщенно сводится к основным типам: атрибутивным, объектным, субъектным, обстоятельственным и комплетивным [23].

Формальные модели атрибутивных отношений представляются в виде:

a) $Atr_{\cap N}\,N$ – модель существительного ($N$) с согласованным атрибутом $Atr$ (согласование по всем грамматическим признакам), например, *конечная группа.*

b) $N{+}Atr$ – модель существительного ($N$) с несогласованным признаком $Atr$ (синтаксическое примыкание), например, *группа без кручения, группа с обычной арифметической операцией умножения.*

Объектные отношения семантически ограничены: главное слово в них обозначает действие, состояние, восприятие, чувствование, а зависимое - объект этого действия, восприятия, чувствования (преимущественно в позиции прямого объекта). Имена действия сохраняют актантную структуру исходного глагола (*решить уравнение – решение уравнения*). Частные модели объектных отношений опираются на модели управления глаголов.

Формальная модель объектного отношения представляется в виде $N_v{+}N_{p2}$, где $N_v$ - отглагольное существительное, $N_{p2}$ – существительное в родительном падеже.

Субъектные отношения характеризуют словосочетания с глаголами в форме страдательного залога и страдательных причастий. Зависимая форма имени существительного в таких словосочетаниях обозначает действующее лицо или предмет (творительный падеж). Например: *предложенный автором* (*метод*). Формальная модель субъектного отношения представляется в виде $V^2/A_v^2{+}N_{p5}$, где $V^2/A_v^2$ - глагол или причастие в страдательном залоге, $N_{p5}$ – сущ. в тв. падеже.

Обстоятельственные отношения (определительно-обстоятельственные, временные, пространственные, причинные, целевые) свойственны глагольным словосочетаниям и опираются на лексическое значение процессуальности, например, *группа параллельных переносов в линейном пространстве* - пространственное отношение, для сравнения *строка в таблице* (*атрибутивная модель*). Формальная модель обстоятельственного отношения представляется в виде $N_v{+}PP$, где $N_v$ - отглагольное существительное, $PP$ – предложная группа. Комплетивные (восполняющие) отношения возникают в устойчивых словосочетаниях (фразеологических моделях). Формальная модель комплетивного отношения представляется списком соответствующих словосочетаний.

В общем случае, многословный термин может быть построен как суперпозиция указанных выше отношений. Например, словосочетания *абелева группа без кручения первого ранга* есть суперпозиция *(((абелева группа) без кручения) первого ранга), (группа (параллельных [переносов в линейном пространстве]))*, где круглые скобки выделяют атрибутивные отношения, а квадратные – обстоятельственные отношения.

Семантическая структура многословных терминологических словосочетаний может быть отражена в структуре терминологической аннотации, в которой выделяются тип связи, главное и зависимые слова. Взаимоотношения составляющих оформляются по типу подчинительной синтаксической связи. Классами терминологической схемы являются класс SyntacticLink (синтаксическое отношение) и класс Term (терминологическое словосочетание), отношения терминологической схемы представлены в Табл. 2.

Терминологическая аннотация использует следующие составляющие:

1. Тип синтаксической связи (SyntacticLink)

Экземплярами класса SyntacticLink являются AttributiveLink (атрибутивное отношение); ObjectiveLink (объектное отношение); SubjectiveLink (субъектное отношение); AdverbialLink (обстоятельственное отношение); CompletiveLink (комплетивное отношение).

2. Структурные составляющие (Holder, Dep) словосочетания (Term)

Term *containsHolder* Holder // Holder - главное слово терминологического словосочетания.

Term *containsDependent* Dep // Dep - зависимое слово терминологического словосочетания.

Пример терминологической аннотации для терминологического словосочетания:
*абелева группа* (= term1): *<term1> formedBy Attributive_Link* .
*<term1> containsHolder "группа".*
*<term1> containsDependent "абелева".*

Информация о структуре терминологического словосочетания позволяет более точно различать семантические контексты, имеющие близкий лексический состав. Терминологическое аннотирование в приложениях к информационному поиску позволяет не только построить терминологический индекс документа, но и более точно выделять релевантные документы. Ожидается, что релевантные запросу документы должны содержать словосочетания, имеющие ту же синтаксическую структуру, что и словосочетания запроса. Например, если запрос содержит словосочетание "*ранг абелевой группы*", то документы, содержащие словосочетание "*абелева группа без кручения первого ранга*" должны быть признаны нерелевантными запросу, так как в словосочетании "*ранг абелевой группы*" термин *ранг* является главным словом (Holder), а в конструкции "*абелева группа без кручения первого ранга*" этот термин находится в зависимой позиции (Dep) в составе атрибутивного отношения.

Для связи с терминологическими словарями вводится отношение *mapsTo* между экземплярами классов Term и элементами skos:Concept, представленными в онтологии контролируемых словарей SKOS [24].

В частности, с помощью введенного отношения можно выразить тот факт, что конкретное терминологическое словосочетание (term1) содержит упоминание концепта "нильпотентная группа" из набора данных DBPedia (dbpedia:Nilpotent_group – 2 вхождения): example. doc#term1 mapsTo dbpedia:Nilpotetnt_group.

## 1.3. Расширенная формульная аннотация

Расширенная формульная аннотация дополнительно включает взаимные связи между описаниями переменных и математическими выражениями, использующими эти переменные. Предполагается, что математический документ содержит специальные текстовые сегменты, задающие описания использованных переменных (например, сегменты введения обозначений и описания семантики переменных), которые удаленно расположены по отношению к соответствующим математическим формулам. Для задания расширенного формата формульной аннотации необходимо установить соответствие между текстовым сегментом, содержащим описания переменных, и сегментом, содержащим формулу. Формально вводятся класс Symbol и семантическое отношение

Табл. 2. Отношения терминологической схемы

| Отношение | Описание | Домен | Диапазон |
|---|---|---|---|
| formedBy | образован посредством | Term | SyntacticLink |
| containsHolder | содержит главное слово | Term | rdfs:Literal |
| containsDependent | содержит зависимое слово | Term | rdfs:Literal |
| mapsTo | отображается на | Term | skos:Concept |
| containsTerm | содержит терминологическое словосочетание | DocumentSegment | Term |

hasTextDefinition ассоциации переменной с ее текстовым определением (например, «пусть μ – *среднее значение*»). Также специфицируется отношение containsSymbol между сегментом и символьным обозначением (например, раздел или формула содержит упоминание символа μ).

### 1.4. Интегрированная семантическая аннотация математического документа

Рассмотренные разметки математического документа имеют общий формат, основанный на представлении метаданных в виде RDF триплетов и словарей на языке OWL. Для того чтобы проиллюстрировать их интеграцию, рассмотрим следующий пример (Рис.2).

На Рис. 2 приведен фрагмент математической научной публикации, содержащий три сегмента – теорему, следствие теоремы и доказательство следствия. Кроме того, в тексте теоремы встречается текстовое определение значения символа G («Пусть $G$ – *конечная группа*»), а следствие содержит формулу с его использованием ($G=K\times H$) и наименование термина *нильпотентная группа*.

На Рис. 3 изображен семантический граф, отражающий элементы всех разметок для данного примера. Вершины графа с овальным контуром, имеющие имена с префиксом example.doc, – сегменты документа example.doc. Вершины с квадратным контуром – концепты, определенные выше, в том числе из набора данных DBPedia (“нильпотентная группа”). Направленные ребра – отношения между элементами.

## 2. Метод терминологического аннотирования на основе онтологических ресурсов

Основной проблемой для эффективного терминологического аннотирования документов является отсутствие общедоступных больших терминологических ресурсов по различным предметным областям, на основе которых можно было бы производить аннотирование. Альтернативным решением является автоматическое извлечение терминологии из текстов. В настоящее время можно выделить три основные

**Теорема 1.** *Пусть $G$ — конечная группа, в которой для любой подгруппы $A$ выполняется неравенство $(*)$. Тогда $G = K\lambda H$, где $H$ — силовская 2-подгруппа группы $G$, подгруппа $K$ абелева, $|H/C_H(K)| \le 2$ и либо группа $H$ абелева, либо $H/Z(H)$ — четверная или диэдральная группа.*

**Следствие.** Если $G$ — неабелева нильпотентная группа, в которой условие $(*)$ выполняется для любой неабелевой подгруппы $A$, то $G = K \times H$, где $H$ — неабелева силовская $p$-подгруппа, строение которой определено в теореме 3, а группа $K$ абелева.

**Доказательство.** Из леммы и теоремы 1 следует, что в группе $G$ только одна силовская подгруппа может быть неабелевой. □

Рис. 2. Фрагмент математической публикации

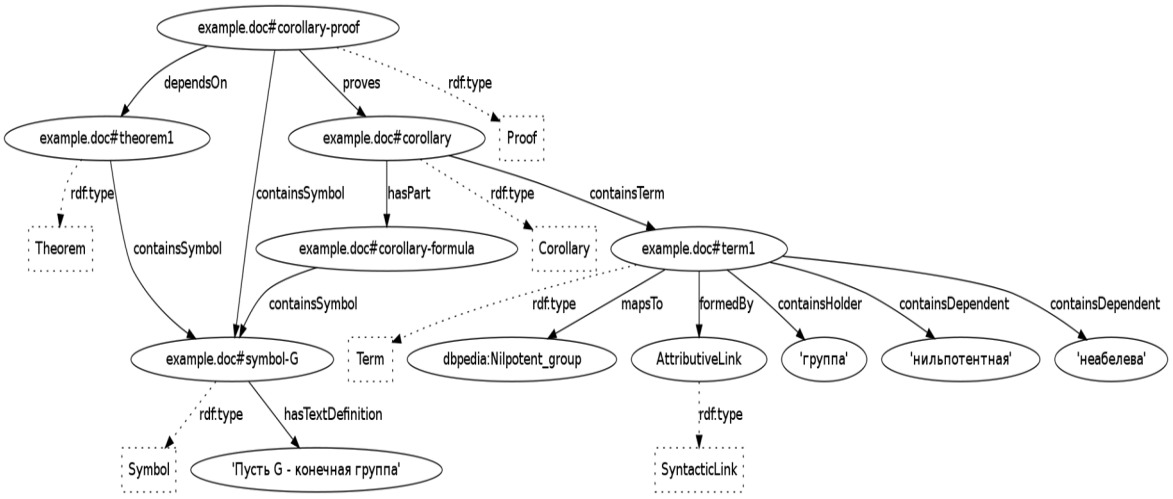


Рис. 3. Семантический граф фрагмента математического документа

группы методов извлечения терминологии из текстов: лингвистические методы, статистические методы и комбинированные методы.

Лингвистические методы базируются на лексико-синтаксических шаблонах однословных и многословных терминов и системе фильтров для отсеивания нетерминов. Применение статистических методов опирается на представление о частотности терминов. Терминосочетания обычно соотносятся с *n*-граммами (двух-, трех-, четырехчленными сочетаниями), характеризующимися высокой степенью устойчивости. В качестве мер, пригодных для оценки устойчивости словосочетаний в текстах, обычно используются *MI-score, t-score, Log-Likelihood, C-value*, критерий $\chi^2$ и ряд других. Комбинированные методы анализа терминологии предполагают совместное использование аппарата лексико-грамматических шаблонов, методов сборки терминосочетаний, системы фильтров, а так же статистического аппарата [25].

Для терминологического аннотирования математического документа нами разработан метод аннотирования на основе онтологических ресурсов (контролируемых словарей).

Выделение в тексте терминологических словосочетаний выполняется с помощью правил синтаксического анализа именных групп. Структурные модели именных групп различаются типом синтаксических отношений между хозяином группы (существительным) и зависимыми членами (модификаторами), при этом допускается произвольное число модификаторов и комбинация любых типов синтаксических отношений (атрибутивные, объектные, субъектные и обстоятельственные) в структуре именной группы.

Разработанный метод позволяет дополнительно выделять именные группы в конструкциях с сочинительным сокращением. При терминологическом анализе сочиненных синтаксических конструкций определенных типов решается обратная задача выделения потенциальных составляющих конструкции для их распознавания как самостоятельных терминологических единиц. Например, в синтаксической конструкции "*операции сложения и умножения натуральных чисел*" выделяются составляющие "*операция сложения натуральных чисел*" и "*операция умножения натуральных чисел*", которые распознаются как отдельные терминологические единицы. Выделение составляющих из сочинительных конструкций производится на основе специальных правил, которые учитывают явление "семантической однородности". Правила выделения составляющих из сочинительных конструкций подробно описаны в [26].

Эксперименты по распознаванию терминологии в тексте на основе основных синтаксических моделей именных групп и моделей сочинительных сокращений были выполнены в онтолингвистической системе "OntoIntegrator" [27] на основе прикладного онтологического ресурса - онтологии математического знания (раздел "теория групп") из DBPedia. Аннотированию документа предшествовала его предобработка, которая включала сегментацию текста на предложения, распознавание объектов текста (выделение формул, числовых последовательностей, слов, знаков препинания, аббревиатур и др.), распознавание именных групп, в том числе содержащих термины прикладной онтологии, распознавание сложных синтаксических конструкций (групп сочинительного сокращения) и другие процедуры (например, выделение и классификация омонимов).

Необходимо отметить, что математические тексты содержат большое число слов с формульно-префиксными частями (например, *p-подгруппа, 2-подгруппа*), а также формульно-постфиксными частями (например, группа *G, подгруппа K*)). В качестве префикса могут быть использованы произвольные формулы и выражения. Такие объекты не содержатся в словаре системы и обрабатываются специальными методами, которые отсекают левый формульный префикс и работают по синтаксической модели правой части слова. Обработка слов с формульно-постфиксными частями производится по синтаксической модели именной группы с аббревиатурой (*группа G, подгруппа K*).

## 3. Метод структурного аннотирования математических научных публикаций в формате LaTeX

Метод структурного аннотирования математических научных публикаций в формате LaTeX детально описан в работе [28]. Предлагаемый метод разбивает задачу структурного аннотирования на две подзадачи: определение (классификация) типов сегментов документа в смысле онтологии; классификация семантиче-

ских отношений между сегментами. При этом качество решения первой задачи оценивается на уровне 89-100% по F1-мере в зависимости от типа сегмента (F1=2*P*R/(P+R), где P – точность, R – полнота), в то время как более сложная задача извлечения и классификации отношений решается на существенно более низком уровне 61-74% по F1-мере.

Классификация типов сегмента документа состоит в следующем. Дано C — множество, как объединение множества классов онтологии Mocassin и множества классов онтологии SALT, S_d — множество сегментов документа d. Необходимо построить отображение O: S_d → C, которое состоит из пар вида (s, c_s), где s — элемент множества S_d и c_s - ассоциированный с ним элемент множества C. Для решения этой подзадачи предлагается следующий алгоритм.

```
1.  for s in S_d do
2.      string:= ExtractSegmentTitle(s)
3.      if string = NIL then
4.          string:= GetEnvironmentName(s)
5.      end if
6.      for c in C do
7.          name:= GetClassName(c)
8.          (c_i, v_i) :=
        ComputeStringSimilarit (string, name)
9.      end for
10.     (c_max, v_max):= GetMaxValue((c, v))
11.     if v_max >= threshold
12.         c_s := c_max
13.     else
14.         c_s:= NIL
15.     end if
16.     add(O, (s, c_s))
17. end for
```

На вход алгоритма подаются множество S_d сегментов документа d, множество C классов онтологии и параметр threshold, значение которого варьируется от 0 до 1. В строках 2 и 4 алгоритма анализируется LaTeX разметка исходного документа. Названия сегментов извлекаются из объявлений вида «\newtheorem{thm}{Теорема}» в заголовке документа. Если эти элементы недоступны (например, вынесены в отдельный стилевой файл), используются имена LaTeX сред (environments). Примеры сред доказательств – «\begin{proof}», «\begin{pf}», «\begin{prf}». Далее подбирается наиболее похожий класс онтологии для ассоциированного строкового значения на основе алгоритма близости строк. Причем функция GetClassName принимает во внимание синонимию имен для концептов (например, «утверждение» или «требование»). Для фильтрации команд, заведомо не имеющих семантики, определенной в онтологии (вроде вспомогательных команд форматирования center, itemize, enumerate), проводится проверка значения алгоритма близости строк с помощью параметра *threshold*.

Оценка алгоритма проводилась на упомянутой тестовой коллекции статей из arXiv.org (Табл. 3).

Оказалось, что многие исходные документы имеют корректно оформленные заголовки, поэтому приведенные высокие результаты не отражают качество работы алгоритма для сложных случаев, когда доступны только объявления сред. При общей высокой правильности результатов по F1-мере, алгоритм неверно разрешает редкие ситуации, связанные с включением в название сегментов дополнительных фраз (например, *Smirnov's Theorem* или *Thurston's Virtual Fibration Conjecture*). В целом, можно утверждать, что подзадача классификации сегментов математической научной статьи решается с высоким уровнем точности и полноты.

Задача извлечения семантических отношений между сегментами документа формулируется следующим образом. Даны: множество R отношений онтологии Mocassin, множество S_d сегментов документа d. Необходимо определить отображение P: S_d×S_d → R между всевозможными парами сегментов документа d и элементами множества R.

Извлечение семантических отношений – традиционно сложная задача из области анализа

Табл. 3. Оценка алгоритма классификации сегментов документа

| Класс | Количество экземпляров | F1-мера |
|---|---|---|
| Axiom | 5 | 1 |
| Claim | 114 | 0.987 |
| Conjecture | 152 | 0.987 |
| Corollary | 1715 | 0.995 |
| Definition | 1838 | 1 |
| Equation | 71314 | 1 |
| Example | 771 | 0.999 |
| Lemma | 4061 | 0.998 |
| Proof | 4943 | 0.997 |
| Proposition | 3052 | 0.999 |
| Remark | 2114 | 1 |
| Theorem | 4670 | 0.991 |
| NIL | 671 | 0.892 |

текстов на естественном языке. В общем случае, ее решение требует: фактического понимания семантики объектов, вовлеченных в отношение; а также поиска экземпляров отношения между данными объектами. Первое замечание накладывает ограничения на множество допустимых отношений между объектами, определенное в онтологии, второе связано с контекстом отношений. В рассматриваемой задаче вводятся следующие ограничения, модифицирующие изначальную постановку задачи:

- *навигационные отношения* онтологии (*refersTo и dependsOn*) должны быть выражены в тексте посредством ссылочных предложений;
- *ограниченные отношения* онтологии (*proves, hasConsequence, exemplifies*) возникают в паре последовательных сегментов.

Таким образом, мы заранее ограничиваем пространство перебора пар сегментов, требуемого в постановке задачи.

Поиск отношений указанных типов реализуется двумя отдельными методами, ниже будет подробно рассмотрен алгоритм извлечения навигационных отношений:

```
1.  for s in S_d do
2.      b_s := GetSegmentBeginning(s)
3.      e_s := GetSegmentEnd(s)
4.      F_s := ExtractReferentialSentences(s)
5.      for f in F_s do
6.          m := FindMentionedSegment(f, S_d)
7.          b_m := GetSegmentBeginning(m)
8.          e_m := GetSegmentEnd(m)
9.          d_1 := b_s — b_m
10.         d_2 := e_s — e_m
11.         verbs := ExtractVerbTokens(f)
12.         add(P, (s, m, d_1, d_2, verbs))
13.     end for
14. end for
```

Вход для алгоритма: множество S_d сегментов документа d. В строках 2 и 3 алгоритм извлекает для каждого сегмента координаты начальной и конечной позиций. В строке 4 извлекаются ссылочные предложения: функция ExtractReferentialSentences выделяет из текста сегмента предложения и находит те из них, которые содержат LaTeX команду ref (например, вида «\ref{thm3.1}»). Для простоты изложения допустим, что каждое ссылочное предложение содержит только одну команду. Тогда для каждого ссылочного предложения находится сегмент, на который оно ссылается. Функция FindMentionedSegment (строка 6) находит или сам сегмент документа по LaTeX команде label (например, теорему с меткой «\label{thm3.1}»), или среду, содержащуюся в сегменте. В последнем случае функция ассоциирует с меткой наименьший объемлющий сегмент. В строках 7 и 8 извлекаются начальная и конечная позиция найденного сегмента. В строках 9 и 10 вычисляется характеристика относительного расположения для данной пары сегментов как разности начальных и конечных позиций, соответственно (характеристика «расположение»). Эта характеристика принимает во внимание не только порядок расположения сегментов, но и их удаленность друг от друга в документе. В строке 11 из ссылочного предложения извлекаются глагольные фразы. Функция ExtractVerbTokens производит разбиение на токены, лемматизацию и анализ частей речи. После этого информация об искомых частях глагольных фраз может быть закодирована в виде обычного булевского вектора по всем словам из глагольных фраз в данной коллекции документов (характеристика «глаголы»). Например,

| | будет | использован | показан | получен |
|---|---|---|---|---|
| «…используя Лемму 1.1 получим …» | 0 | 1 | 0 | 1 |
| «…как будет показано в разделе 2…» | 1 | 0 | 1 | 0 |

Таким образом, выходные данные алгоритма включают три типа характеристик: «расположение», «глаголы» и типы сегментов. Для экспериментов были случайным образом отобраны 243 ссылочных предложения из разных документов, для которых вручную были установлены типы семантических отношений и автоматически вычислялись значения трех указанных выше характеристик. На основе полученных значений строился классификатор на базе деревьев принятия решений (алгоритм C4.5 [29]). Эксперимент показал, что 95% ссылочных предложений действительно сгенерировали отношения типа refersTo или dependsOn (остальные случаи помечались как other и включали экземпляры отношения proves и других). Стандартные оценки корректности работы метода классификации представлены в Табл. 4.

Результаты показывают, что имея небольшой набор размеченных отношений, можно корректно классифицировать до 75% от общего

Табл. 4. Оценка алгоритма классификации навигационных отношений

| Характеристики | Правильность | F1-мера для refersTo | F1-мера для dependsOn |
|---|---|---|---|
| типы | 0.663 | 0.566 | 0.752 |
| типы+расположение | 0.658 | 0.663 | 0.704 |
| типы+глаголы | 0.704 | 0.653 | 0.770 |
| все | **0.741** | **0.744** | **0.772** |

числа ссылочных предложений, устанавливая, таким образом, семантические отношения для сегментов, вовлеченных в них. Также заслуживает отдельного исследования распределение значений характеристик в объеме большего корпуса примеров (например, по всему объему имеющихся тестовых коллекций) для улучшения производительности классификации. Обнаружение возможных полезных закономерностей в этих данных могло бы позволить комбинировать их с данными из небольшого набора размеченных вручную отношений.

## 4. Семантический поиск в размеченных математических коллекциях

Поисковые механизмы, разработанные на основе семантически размеченных математических коллекций, позволяют адекватно выполнять сложные поисковые запросы, которые не могут быть выполнены на основе стандартного поиска по ключевым словам.

Проблема поиска по структуре математических документов формулируется в терминах фактологического поиска в некоторой базе знаний. В рамках предлагаемого подхода база знаний представляет собой RDF-хранилище, которое содержит проаннотированные объекты, извлеченные из исходных документов, т.е. элементы логической структуры, терминологические и формульные объекты и их отношения.

В настоящем разделе рассматривается классификация и формальная интерпретация поисковых запросов, выполняемых к базе знаний объектов семантических аннотаций. Для формализации запросов выбран язык SPARQL как де-факто стандарт языка запросов к RDF графам. Многие современные RDF-хранилища (в частности, Virtuoso) реализуют расширенные версии SPARQL-процессоров, позволяя, например, комбинировать семантические запросы с обычными запросами по ключевым словам. На данном этапе исследований авторами выделяются три типа запросов к RDF-хранилищу, которые рассматриваются как новые реализуемые типы, в том смысле, что данные типы запросов не исполняются удовлетворительно современными поисковыми системами.

Классификация конкретизирует каждый тип поисковых запросов по характеру объектов, использованных в запросе, а также с точки зрения применения определенных терминологических словарей.

Ниже приводятся классификация и формальная интерпретация новых типов запросов, выполняемых на базе семантических аннотаций, их содержательное описание, примеры, а также интерпретация на языке SPARQL. Элементы с префиксом o: суть элементы интегрированной схемы всех трех представленных разметок.

- Поиск структурных элементов на основе онтологии структуры математического документа в качестве терминологии.

Пример поискового запроса данного типа на естественном языке - «*найти доказательства теорем*». На языке SPARQL данный запрос формулируется следующим образом:

*SELECT ?p WHERE { ?proof a o:Proof .*
*?proof o:proves o:Theorem }*

При формализации запроса используются только элементы структурной разметки.

- Поиск структурных элементов и объектов математического знания.

Пример – «*найти теоремы, упоминающие термины из теории групп*». В этом случае осуществляется поиск структурных элементов - теорем, имеющих отношение к конкретной области математического знания. Для формализации запроса используются онтология структуры математического документа и терминологические словари. На языке SPARQL данный запрос формулируется следующим образом:

*SELECT ?t WHERE {?theorem a o:Theorem .*
*?theorem containsTerm ?term .*
*?term mapsTo ?concept .*
*?concept skos:subject dbpedia:Group_theory }*

В этом запросе используются элементы структурной и терминологической разметки.

- Поиск структурных элементов по символьным обозначениям математических терминов.

Данный тип запросов расширяет возможности стандартного поиска по формулам, так как позволяет использовать связь между математическими терминами и их символьными обозначениями в тексте. Пример – «найти формулы, в которых встречается символ, определенный для обозначения конечной группы». Запрос на языке SPARQL:

*SELECT ?equation WHERE {?equation a o:Equation .*
*?equation o:containsSymbol ?symbol .*
*?symbol o:hasTextDefinition ?textDefinition .*
*?textDefinition bif:contains "конечная AND группа" }*

В запросе используются элементы структурной и формульной разметок, а также свойство bif:contains расширенного SPARQL-процессора для комбинирования с поиском по ключевым словам.

## Заключение

В статье предложена модель аннотации математического текста, включающая различные семантические компоненты. Структурная аннотация математического документа описывается онтологией структурных элементов математической научной публикации. Терминологическая аннотация использует набор контролируемых словарей, для которых разрабатываются методы их автоматического пополнения. Расширенная формульная аннотация позволяет формировать контекст формулы, включающий все релевантные текстовые фрагменты и размеченные формулы, в том числе, фрагменты обозначения переменных формулы. Предложенная модель, опираясь на стандартные средства Семантического Веба, позволяет в едином формате представления данных связать между собой различные семантические аннотации. Таким образом, на основе интегральной модели документа можно построить семантически размеченную коллекцию математических текстов, которую можно использовать в различных прикладных задачах.

Следующий шаг исследований – разработка эффективных автоматических методов аннотирования математических коллекций. В этом направлении уже начаты работы, некоторые результаты которых изложены в настоящей статье. Автоматическое аннотирование включает методы структурного и терминологического аннотирования, а также методы автоматической сборки расширенного контекста формулы. В данной статье рассматривались методы структурного аннотирования LaTeX –документов и метод терминологического аннотирования на основе онтологий. Метод терминологического аннотирования базируется на синтаксических моделях именных групп в русском языке и позволяет выделять в текстах синтаксические конструкции произвольной сложности.

Дальнейшее развитие представленного подхода связано с разработкой методов автоматического аннотирования расширенного формульного контекста, метода структурного аннотирования математических текстов в формате PDF, а также методов автоматического извлечения терминологии, используя, в частности, методы машинного обучения.

Разработанную модель аннотации математического документа предполагается использовать в различных приложениях. Одним из важных направлений является использование семантически аннотированной коллекции математических текстов для улучшения качества математического поиска. В статье приведены новые типы запросов, которые не могут быть удовлетворительно выполнены современными поисковыми машинами, но эффективно выполняются на основе предложенного подхода. Можно указать ряд важных задач, в которых предлагается использовать семантически аннотированные математические коллекции: аннотирование и поиск наиболее важных результатов математических статей, расширенный поиск по формулам, классификация документов, извлечение специальных математических объектов из текстов в целях реферирования и др.

## Литература

1. Хорошевский В.Ф. Пространства знаний в сети Интернет и semantic Web (Часть 1) // Искусственный интеллект и принятие решений. – 2008. – № 1. – С. 80-97.
2. Youssef, A. Roles of Math Search in Mathematics // Mathematical Knowledge Management, 5th International Conference, MKM 2006, Wokingham, UK, August 11-12, Proceedings. – Berlin: Springer, 2006. – P. 2 – 16.
3. CiteSeerX [Электронный ресурс]. URL: http://citeseerx.ist.psu.edu.
4. Google Scholar [Электронный ресурс]. URL: http://scholar.google.com.

5. Math-Net.Ru [Электронный ресурс]. URL: http://www.mathnet.ru.
6. Zentralblatt MATH [Электронный ресурс]. URL: http://www.zentralblatt-math.org/zmath.
7. arXiv [Электронный ресурс]. URL: http://arxiv.org.
8. Kohlhase M, Sucan I. A search engine for mathematical formulae // Artificial intelligence and symbolic computation, 8th international conference, AISC 2006, Beijing, China, September 20 – 22, 2006. Proceedings. – Berlin: Springer, 2006. – Lecture Notes in Computer Science. – Vol. 4120. – Lecture Notes in Artificial Intelligence. – P. 241–253.
9. Miner R., Munavalli R. An approach to mathematical search through query formulation and data normalization // Towards mechanized mathematical assistants. 14th symposium, Calculemus 2007, 6th international conference, MKM 2007, Hagenberg, Austria, June 27 – 30, 2007. Proceedings.– Lecture Notes in Computer Science. – Vol. 4573. Lecture Notes in Artificial Intelligence. – Berlin: Springer, 2007. – P. 342 – 355.
10. The Wolfram functions site [Электронный ресурс]. URL: http://functions. wolfram.com.
11. Bancerek G., Rudnicki P. Information Retrieval in MML // Mathematical Knowledge Management, 3rd International Conference, MKM 2003, Proceedings. – Lecture Notes in Computer Science. – Springer, 2003 – Vol. 2594. – P. 119 – 132.
12. Cairns P. Informalising Formal Mathematics: searching the mizar library with Latent Semantics // Mathematical Knowledge Management. – Springer, 2004. – P. 58 – 72.
13. de Waard, A. and Tel, G. The ABCDE Format – Enabling Semantic Conference Proceedings // Proceedings of the 1st Workshop on Semantic Wikis, European Semantic Web Conference, ESWC 2006, Budva, Montenegro, 2006 [Электронный ресурс]. URL: http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-206/paper8.pdf.
14. Groza T. et al. SALT – semantically annotated LaTeX for scientific publications // Proceedings of the 4th European Semantic Web Conference, ESWC 2007. – Lecture Notes in Computer Science. – Springer, 2007. – Vol. 4519. – P.518 – 532.
15. Nakagawa K., Nomura A., Suzuki M. Extraction of logical structure from articles in mathematics // Mathematical Knowledge Management. – Lecture Notes in Computer Science. – 2004. – Vol. 3119. – P. 276 – 289.
16. Asperti A. et al. Mathematical knowledge management in HELM // Annals of Mathematics and Artificial Intelligence. – 2003. – V. 38. - No 1-3. – P. 26 – 46.
17. Schena I. Towards a Semantic Web for Formal Mathematics // Technical Report UBLCS. – 2002 [Электронный ресурс]. URL: http://www.cs.unibo.it/pub/TR/UBLCS/2002/2002-06.ps.gz.
18. Kohlhase M. OMDoc – An open markup format for mathematical documents [Version 1.2] // Lecture Notes in Artificial Intelligence. – Springer, 2006. – Vol. 4180. – P. 28 – 32.
19. Lange C. Integrated Semantic Web Collaboration on Semiformal Mathematical Knowledge. PhD Thesis. – Jacobs University Bremen, 2011 [Электронный ресурс]. URL: https://svn.kwarc.info/repos/swim/doc/phd/phd.pdf.
20. Жильцов Н.Г. Онтология структуры публикаций по математике // Системный анализ и семиотическое моделирование: материалы первой всероссийской научной конференции с международным участием. – Казань: Фэн, 2011. – С. 46 – 52.
21. DBPedia [Электронный ресурс]. URL: http://dbpedia.org.
22. Mathematical thesaurus [Электронный ресурс]. URL: http://thesaurus.maths.org.
23. Валгина Н.С. Синтаксис современного русского языка. – М.: Агар, 2000. – 416 с.
24. SKOS Simple Knowledge Organization System [Электронный ресурс]. URL: http://www.w3.org/2004/02/skos/
25. Лукашевич Н.В., Логачев А.М. Комбинирование признаков для автоматического извлечения терминов // Вычислительные методы и программирование. – 2010. – Т. 11.– С. 108 – 116 [Электронный ресурс]. URL: http://num-meth.srcc.msu.su/.
26. Невзорова О.А., Невзоров В.Н. Онтологический анализ прикладной области: автоматизированные методы выделения терминов в системе "OntoIntegrator" // Методы моделирования. – Казань: Фэн, 2009. – С. 196 – 208.
27. Nevzorova O., Nevzorov V. The development support system "Ontointegrator" for linguistic applications // Intelligent Information and Engineering Systems. – 2009. – No. 13.– P. 78 – 84.
28. Solovyev V., Zhiltsov N. Logical Structure Analysis of Scientific Publications in Mathematics // Proceedings of the International Conference on Web Intelligence, Mining and Semantics, WIMS 2011, Sogndal, Norway, May 25 – 27, 2011. – ACM, 2011. – P. 21:1 – 21:9.
29. Kamareddine F., Wells J.B. Computerizing Mathematical Text with MathLang // Electronic Notes in Theoretical Computer Science. – 2008. – Vol. 205. – P. 5– 30.

**Невзорова Ольга Авенировна**. Заместитель директора НИИ «Прикладная семиотика» АН РТ (Казань). Окончила Казанский государственный университет в 1982 году. Кандидат технических наук. Автор 98 печатных работ. Область научных интересов: искусственный интеллект, онтологическое моделирование, компьютерная лингвистика. E-mail:onevzoro@gmail.com

**Биряльцев Евгений Васильевич.** Заведующий лабораторией НИИ математики и механики им. Н.Г. Чеботарева Казанского (Приволжского) федерального университета. Окончил Казанский государственный университет в 1983 году. Кандидат технических наук. Автор 27 печатных работ. Область научных интересов: суперкомпьютерные вычисления, архитектура программных систем, математическое моделирование. E-mail: IgenBir@yandex.ru

**Жильцов Никита Геннадьевич.** Аспирант Казанского (Приволжского) федерального университета. Окончил Казанский государственный университет в 2009 году. Автор 6 печатных работ. Область научных интересов: семантический поиск, семантические технологии обработки текстов. E-mail: nikita.zhiltsov@gmail.com