OSTIS-2012
(Open Semantic Technologies for Intelligent Systems)

УДК 004.9:510

ПОДХОД К СЕМАНТИЧЕСКОМУ ПОИСКУ МАТЕМАТИЧЕСКИХ ВЫРАЖЕНИЙ В НАУЧНЫХ ТЕКСТАХ

Биряльцев Е.В.[*], Галимов М.Р.[*], Жильцов Н.Г.[*], Невзорова О.А. [**]

[*] *Казанский (Приволжский) федеральный университет, г. Казань, Россия*

IgenBir@yandex.ru

glmvmrt@gmail.com

nikita.zhiltsov@gmail.com

** *Научно-исследовательский институт «Прикладная семиотика» АН Республики Татарстан, г. Казань, Россия*

onevzoro@gmail.com

В работе рассматривается подход к семантическому поиску математических выражений, позволяющий выполнять запросы на поиск математических формул по текстовым наименованиям переменных, входящих в формулы. Обсуждается прототип системы поиска и методы решения основных проблем предложенного подхода.

Ключевые слова: семантический поиск, математический поиск, разметка формул.

Введение

Поиск по математическим документам [1] – актуальная и быстроразвивающаяся область исследований. Современные математические поисковые системы условно можно разделить на три группы. К первой относятся системы поиска научных публикаций и поисковые интерфейсы крупнейших научных коллекций, 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. К числу таких систем относятся хорошо известные системы, такие как Google Scholar [Google], Microsoft Academic Search [Microsoft], CiteseerX и Scirus. Полнотекстовый поиск по ключевым словам достаточно удобен для конечного пользователя.

Отличительная особенность систем второй группы состоит в том, что они используют семантику математической нотации и реализуют поиск по формулам и выражениям. Данные системы работают со специальным семантическим представлением математических формул, выраженным на языках Content/Presentation MathML и OpenMath. В качестве результатов поиска возвращаются ссылки на документы, содержащие релевантные формулы. Основными трудностями при разрешении запросов данного типа являются: зависимость форм нотации от контекста; неоднозначность трактовки одних и тех же символов; идентичность формул с точностью до обозначений. Специализированные системы поиска по математическим формулам предлагают средства для формулирования запроса в синтаксисе языка разметки LaTeX (например, Springer LaTeXSearch [LatexSearch] индексирует базу статей издательства Springer; (uni)quation [Uniquation] — научные тематические сайты, форумы и Wikipedia) или MathML, используя соответствующие графические интерфейсы [MathWebSearch].

Третья группа – приложения], в основном, представляющие собой расширения систем автоматических доказательств, использующие строго формализованное представление математических документов на таких языках, как Mizar и Coq. 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.

Подход, представленный в настоящей статье, направлен на интеграцию функциональных возможностей полнотекстового поиска и поиска по математическим формулам, при котором конечному пользователю предлагается формулировать поисковый запрос на поиск математической формулы в форме ключевых слов. Наиболее близким подходом является подход математической поисковой системы EgoMath (доступна по адресу http://egomath.projekty.ms.mff.cuni.cz), которая в данный момент предоставляет возможности традиционного формульного поиска в синтаксисе LaTeX и механизм переформулирования запроса (от ключевых слов к символьным обозначениям), однако алгоритм этого связывания не раскрыт в оригинальной статье авторов EgoMath [Misutka et al., 2008]. Новизна предлагаемого подхода состоит в том, что основной метод связывания рассматривает локальный контекст формулы, включающий текстовое расширение и тип структурного элемента (например, теорема, доказательство или раздел), содержащего математическое выражение. Это позволяет, с одной стороны, более точно определять семантику переменных в формуле и, с другой стороны, представлять пользователю релевантный фрагмент документа в выдаваемых результатах. Как правило, возможность выдачи более точно локализуемых положений формулы также не предоставляется поисковыми системами, рассмотренными выше.

Кратко рассмотрим основные подходы к поиску математических выражений.

Известна постановка задачи поиска математических формул по ее фрагментам, например, найти все формулы, в которые входит конструкция "$X^n + 1$" . Данный подход в настоящее время доведен до уровня экспериментальных систем [wolframalpha.com], [http://uniquation.ru/ru/]. Несмотря на внешнюю простоту данной задачи, ее решение в классических полнотекстовых системах затруднено. Если мы хотим указать достаточно сложную математическую конструкцию, то нам необходимо иметь возможность использовать в запросах некоторый язык разметки математических формул. Таким образом, задачу поиска математических формул необходимо решать в рамках контекста некоторой группы языков разметки, причем как на уровне интерфейса пользователя, так и внутренних механизмов индексирования и выполнения запросов.

Для наиболее популярных языков математической разметки TeX и MathML известны решения по конвертации TeX/LaTeX в MathML и частичных обратных преобразований [http://www.w3.org/Math/Software/mathml_software_cat_converters.html]. В рамках настоящей работы мы будем рассматривать только запросы к текстам, содержащим математические формулы в символьном виде с использованием языков математической разметки. При использовании языков разметки презентационного уровня, таких как TeX или MatML презентационного уровня, возникают некоторые проблемы, связанные с семантикой математических формул, в частности, не определяются свойства операций (коммутативность, ассоциативность, транзитивность), что влияет на выполнение эквивалентных формульных преобразований. Для приведенного выше примера семантика операции "+" может допускать или не допускать коммутативность, поэтому выражения "$X^n + 1$" и "$1 + X^n$" могут быть семантически эквивалентными или неэквивалентными. Семантический уровень математический разметки, такой как MathML семантического уровня или OpenMath [http://www.openmath.org/], частично снимает эту проблему. Фиксированный (MathML) или расширяемый (OpenMath) набор математических операций позволяет однозначно описать и разобрать описание сходных (по написанию) математических конструкций. Можно встроить в поисковые механизмы алгоритмы, учитывающие свойства операций, что повышает релевантность ответов на запрос, т.е. в ответах выдаются ссылки на документы, содержащие формулы вида "$X^n + T + 1$", "$1 + X^n$", но не "$N^x + 1$" (коммутативность и ассоциативность операции "+" и некоммутативность операции возведения в степень).

Дальнейшее рассмотрение темы полноты и релевантности поиска математических формул по их фрагментам приводит в необходимости решения вопроса семантической эквивалентности выражений "$X^n + 1$" и, например, "$Y^n + 1$", т.е. вопроса об эквивалентности математических выражений в различных системах обозначений переменных. Здесь необходимо различать, что собственно ищет пользователь - абстрактную математическую конструкцию или в переменные вкладывается некоторый нематематический смысл. Очевидно, что если речь идет о поиске собственно математических конструкций с некоторой структурой заданной последовательностью математических операций, то выражения "$X^n + 1$" и "$Y^n + 1$" эквивалентны. В противном случае необходимо рассматривать нематематическую семантику переменных, участвующих в поисковых запросах.

Данную интерпретацию можно рассматривать как новый тип запросов к коллекциям математических документов. Действительно, в отличие от рассмотренной задачи поиска математической формулы по ее фрагменту в формулировке запроса должны использоваться не математические конструкции, а словесные наименования переменных, входящих в искомую формулу. Результатом поискового запроса также должен быть тематический реферат найденного текста, содержащий искомую формулу и определения обозначений использованных в ней имен переменных. В отличие от полнотекстовых запросов результатом поиска является не просто фрагменты документа из коллекции, содержащие слова строки запроса в определенном количестве и

находящиеся в достаточной близости, а таковые фрагменты, содержащие нетекстовый объект (формулу) и фрагменты документа, связывающие переменные формулы и их текстовые определения вне зависимости от их местонахождения в тексте. Для выполнения поискового запроса нам необходимо выделить в анализируемом тексте обозначение переменных по их текстовому описанию и формулы, в которых данные переменные представлены наиболее полно. Таким образом, данный тип запроса не сводится ни к поиску формул по их фрагментам, ни к полнотекстовому поиску, и требует рассмотрения возможности его реализации, особенно в части методов разметки и индексирования коллекций.

В качестве иллюстрации рассмотрим математическую конструкцию вида $a = bc^2$. Как минимум, две популярные формулы имеют такую структуру: формула вычисления площади круга $S = \pi R^2$, и формула, связывающая массу и полную энергию, $E = mc^2$. Поиск только по математическому контексту осложняется тем, что в случае с естественнонаучными формулами мы не можем просто указать, что требуется найти формулу, содержащую переменные S и R, или E и m, так как многие символы перегружены и имеют в различных областях науки различный смысл. Так S - это площадь, пройденный путь, энтропия, а также символ химического элемента серы; E - символ энергии, напряженности электромагнитного поля, прописная e - основание натуральных логарифмов и т.п. Кроме того, в новых и специальных областях науки и техники устоявшиеся обозначения могут отсутствовать, различные научные школы могут придерживаться различных обозначений. Словесная терминология, как правило, более устойчива, В научных текстах "энергия" и "площадь" однозначно обозначают соответствующие физические величины. Таким образом, для поиска конкретной формулы необходимо задать запрос, в котором явно указывается, что требуется найти формулу, связывающую "радиус круга" и "площадь", или формулу с параметрами "энергия", "масса" и "скорость света".

1. Подход к семантическому поиску математических выражений

В рассматриваемой постановке задачи в естественнонаучных текстах выделяются следующие виды сущностей: естественнонаучные термины, символьные условные обозначения терминов, математические фрагменты (формулы) и структурные элементы документа (теоремы, доказательства и пр.), содержащие формулы. Все перечисленные сущности составляют расширенный контекст формулы.

Таким образом, для решения поставленной задачи необходимо вычислять в тексте расширенный контекст формулы, который включает выделение отношений: «термины - условные обозначения» и «условные обозначения – формулы». Первое отношение есть текстовое определение значения символа в некотором контексте с помощью терминов, второе отношение указывает на вхождение символа в формулу. Пример расширенного формульного контекста приведен на рис. 1, который содержит фрагмент статьи Wikipedia о площади треугольника. Из рисунка видно, что определения переменных a, b, c как длин сторон треугольника и α, β, γ как углов треугольника даны в структурном элементе «примечание», а формула, которая связывает данные переменные, находится в структурном элементе «теорема синусов». Таким образом, локальный контекст формулы включает структурный элемент типа «теорема», а также структурный элемент типа «примечание», в котором содержится спецификация семантики формульных переменных. Эта информация отображается в поисковых результатах, что упрощает задачу пользователя по визуальной оценке релевантности найденной формулы.

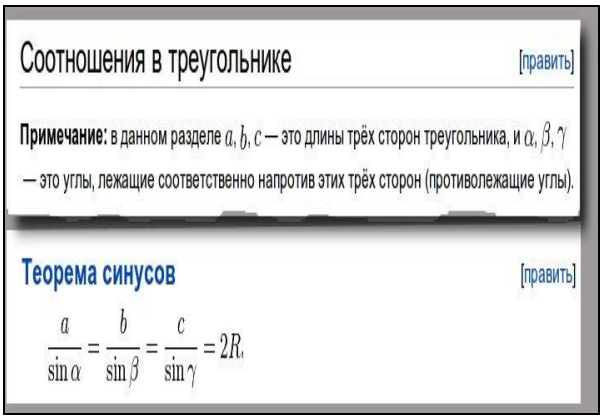

Рисунок 1 – Пример расширенного формульного контекста (фрагмент статьи Wikipedia)

Фиксация выделенных сущностей и зависимостей между ними может представляться как в виде разметки (внесения дополнительных символьных конструкций непосредственно в анализируемый текст), так и построением соответствующих индексов (внешних файлов по отношению к анализируемому тексту). В статье детально рассматривается метод аннотирования, так как построение индексов является более техническим аспектом, связанным с быстродействием, объемами данных и др. аспектами реализации, кроме того индекс всегда может быть построен по разметке. Затем рассматриваются концептуальные аспекты подхода к поиску. При этом технические решения реализация поиска могут существенно различаться в зависимости от объемов коллекции, структуры запросов к ней и других факторов. В частности, на реализацию могут существенно повлиять возможности полнотекстовой поисковой машины, используемой совместно с рассматриваемыми специализированными поисковыми алгоритмами, требования к полноте и релевантности результата,

наличие других разметок и индексов, например разметка логической структуры естественнонаучного текста.

1.1. Разметка и индексирование

Математические выражения в естественнонаучных текстах в настоящее время задаются с использованием специализированных языков разметки (MathML презентационного и семантического уровня; Tex, LaTex, MathType и их конверторы в MathML; OpenMath и др.). Рассмотрим цель дополнительной разметки в рассматриваемой задаче. Как отмечалось выше, связующим звеном между элементами поискового запроса (текстовыми определениями некоторых переменных и результатом поиска в виде математической формулы) является набор условных обозначений данных переменных. Эти условные обозначения представляют собой фрагменты файла с MathML-разметкой. При анализе разметки необходимо выяснить какие из фрагментов являются (или могут являться) математическими формулами, какие - символьными обозначениями переменных, а какие - не относящимися к рассматриваемому вопросу конструкциями, например, таблицами или диаграммами.

Разметка MathML презентационного уровня включает теги, позволяющие представить в математических выражениях все необходимые символы и расположить их в соответствии с правилами записи соответствующей математической нотации. Элементарными конструкциями в MathML презентационного уровня являются токены (символьные последовательности, ограниченные тегами). Например, <mi> - идентификатор; <nm> - число; <ms> - литерал; <mo> - оператор, ограничитель, разделитель; <mtext> - текст (комментарий). Символы внутри тегов могут представляться обычными ASCII-символами, либо шестнадцатеричными кодами Unicode, что позволяет использовать все необходимые специальные символы. Остальные теги позволяют управлять размещением токенов относительно друг друга, в том числе формировать дроби, индексы, надстрочные и подстрочные символы (например, пределы интегрирования), матрицы, и управлять отступами между токенами.

Тег <mi> ограничивает многосимвольные описания идентификаторов, при этом описание сложной переменной (с индексами, подстрочными и надстрочными символами) никак не ограничивается. Таким образом, имена используемых в формулах переменных необходимо вычислять путем сложного контекстного анализа (не всегда однозначного). В MathML презентационного уровня также отсутствует явное указание на то, является ли некоторый фрагмент описанием уравнения или описанием переменной. Можно использовать атрибут *class* тега <mo>, который может принимать значения *open, close, op, rel* для обозначения открывающих и закрывающих скобок различного типа, операций (в том числе имеющих буквенное написание) и отношений. Наличие тега <mo> с атрибутом *class* позволяет сделать вывод, что фрагмент описания не является определением переменной, а представляет более сложную математическую конструкцию.

Если трактовать запросы рассматриваемого типа как поиск уравнений, а не просто математических выражений, то можно еще более сузить класс релевантных запросу объектов, рассматривая только те математические конструкции, в которых присутствует некоторое отношение (как правило, знак равенства). Этот признак также может быть выделен из разметки путем анализа атрибута *class* тега <mo>.

Таким образом, на первом проходе при анализе разметки можно выявить фрагменты размеченного теста, предположительно содержащие имена переменных в окружении сплошного текста и фрагменты, предположительно содержащие некоторые уравнения или отношения.

На втором проходе делается попытка установить связь между "отдельно стоящими" именами переменных и фрагментами - кандидатами в формулы. Данный этап позволяет отсеять отдельные фрагменты, не связанные с формулами, и определить принадлежность переменных формулам. Установление принадлежности можно проводить прямым поиском вхождения выделенных фрагментов (кандидатов в имена переменных) во фрагменты (кандидатов в уравнения). Результатом анализа является взаимная разметка фрагментов установленными связями.

Разметка MathML семантического уровня включает значительно большее количество тегов. Для рассматриваемой задачи важно, что на данном уровне присутствуют теги, позволяющие явно описать сколь угодно сложную переменную (тег <ci>), и тег <apply>, явно указывающий на то, что его содержимое является математическим выражением. В последнем случае, если требуется выделить только выражения, содержащие отношения, в частности, отношение равенства, необходимо анализировать наличие тега <eq> (для выделения уравнений) и тегов других отношений (в более общем случае).

Таким образом, выделение фрагментов, содержащих описания переменных и описания математических выражений (в частном случае, уравнений), при наличии разметки семантического уровня упрощается и становится более определенным. Вместе с тем, это не отменяет выявление связей между переменными и формулами, аналогично соответствующему этапу анализа разметки презентационного уровня. Установление связей между фрагментами (определениями переменных) и фрагментами (формулами) выполняется поиском соответствующих вхождений. Результатом анализа

является взаимная разметка установленными связями.

Более богатый синтаксис MathML семантического уровня ставит некоторые вопросы, отсутствующие на презентационном уровне. Так, наличие тегов <declare> и <lambda>, позволяющих вводить пользовательские типы переменных и функций, требует дополнительного анализа. В простейшем случае можно игнорировать все описания пользовательских типов переменных и функций, так как они описывают математическую семантику вводимых пользователем обозначений. Но, вполне вероятно, что из семантики пользовательских определений можно извлечь дополнительную информацию для решения рассматриваемых задач.

Формализация результатов взаимной разметки может быть произведена в рамках синтаксиса семантического MathML. Для этого можно использовать тег <semantics>, который включает первый атрибут как произвольное высказывание семантического MathML, и произвольное количество последующих атрибутов как произвольные текстовые или xml-последовательности, при этом семантика данных атрибутов не регламентируется MathML, а интерпретация осуществляется обрабатывающими программами. Различать соответствующие семантики позволяет атрибут *encoding* элемента тега <annotation>, в котором указывается имя обрабатывающей программы или иной ключ, позволяющий определять, как интерпретировать содержащуюся в данном атрибуте информацию. Таким образом, в качестве первого атрибута указывается найденный фрагмент текста MathML (формула), а качестве дополнительных атрибутов фиксируются ссылки на соответствующие формулы, задающие описания переменных. Без ограничения общности, на данном этапе рассматривается двусторонняя разметка (от формул к переменным и наоборот), также пока является открытым вопрос о формате записи взаимных ссылок (через смещения относительно друг друга, указания смещения от начала файла или каким другим способом).

Таким образом, построенная аннотация на основе математической разметки позволяет эксплицировать неявную информацию о связях между фрагментами (описаниями переменных) и фрагментами (формулами) и зафиксировать ее в виде взаимной системы ссылок. Процедура разметки включает операцию анализа взаимного вхождения фрагментов с вычислительной сложностью порядка n^2, где n - количество выделенных фрагментов, поэтому выполнение запроса «на лету» (без предварительного аннотирования текста) является весьма трудоемким. Другой важной задачей является задача установления связей между синтаксически тождественными определениями переменных. Действительно, тождественные фрагменты MathML-текста, идентифицированные как описания переменных, могут встречаться в тексте неоднократно. На каждое из таких вхождений, согласно рассматриваемой концепции, требуется зафиксировать ссылку в соответствующей формуле, что многократно увеличивает объем ссылок. В этой ситуации желательно провести некоторую "нормализацию" полученных отношений, составив словарь определений переменных с последующим его использованием для формирования ссылок в формуле. Для решения этой задачи требуется получить семантическую интерпретацию фрагмента-определения переменной в тексте. Определение включает словесное описание некоторой символической конструкции, которая в дальнейшем может многократно использоваться в формулах. В то же время, появление той же символической конструкции вне формулы может означать переопределение ее смысла, а также обсуждение некоторых свойств данной величины или математической абстракции. Априори практически невозможно разделить эти случаи, поэтому создание единого словаря, в котором, возможно, смешиваются случаи вхождения величин, имеющих разный смысл, но одинаковое обозначение, представляется на данном этапе анализа преждевременным.

Рассмотрение вышеуказанных проблем ставит задачу о возможности и целесообразности предварительной семантической разметки. Семантическая разметка предполагает маркирование текстовых фрагментов элементами фиксированного (или расширяемого) словаря (тезауруса, онтологии) и важным моментом является выбор необходимого семантического уровня представления. Так, использование тематических классификаций достаточно затруднительно в связи с ориентацией на широкий класс предметных областей, а применение высокоуровневых онтологий имеет слишком высокий уровень абстракций. Возможным решением является использование онтологии структурных элементов [Solovyev et al., 2010], разработанной авторами статьи в стиле подхода OMDoc [https://trac.omdoc.org/OMDoc].

Важнейшей разработкой последних лет является основанный на XML формат математических документов OMDoc, предназначенный для детализированного представления содержимого математических текстов и впоследствии расширенный на другие разделы науки. Формат OMDoc выделяет три уровня структурной семантической разметки. Первый уровень – уровень объектов. Для этого OMDoc содержит выразительные средства для разметки формул в форматах OpenMath и MathML. Второй уровень – уровень утверждений – для выделения теорем, лемм, определений, ссылок. Наконец, третий уровень – уровень теорий, который предоставляет средства для выражения связей между прикладными математическими теориями. Связанная с форматом OMDoc онтология, реализованная на языке OWL-DL, концептуально выражает терминологию

данного формата и содержит описание структурных элементов математического документа как классов и отношений между ними.

Группой исследователей из Бременского университета разработаны методы для полуавтоматического извлечения структурных элементов из математических текстов. Основная технология выглядит следующим образом. Исходный текст математического документа аннотируется с помощью команд макропакета sTeX [Kohlhase M., 2005], который представляет собой расширение популярного в математическом сообществе пакета AMS-LaTeX. С помощью программной утилиты LaTeXML [Miller B., 2007], которая использует мэппинг между командами sTeX и тэгами OMDoc, исходный текст транслируется в формат OMDoc. Далее из математического документа в формате OMDoc автоматически извлекаются RDF триплеты – утверждения вида субъект-предикат-объект в терминах онтологии OMDoc [Lange C., 2009]. Таким образом, в полуавтоматическом режиме происходит обогащение семантической информации о структуре исходных текстов. Другая технология представления математического документа, развиваемая на базе OMDoc, предполагает представление исходных математических текстов в формате XHTML страниц с семантическими вставками в виде RDFa аннотаций [RDFa, 2008]. Это позволит будущим реализациям семантических поисковиков извлекать дополнительные метаданные непосредственно из страниц, опубликованных в вебе.

Подход OMDoc, предусматривающий разметку высокоуровневых структурных элементов математического текста, таких как теорема, доказательство, определение, пример и др., и связей между ними, позволяет наметить решение указанной выше проблемы разрешения смысла многократных вхождений переменной в сплошной текст. Математическая культура диктует употребление переменных в рамках рассмотрения одного вопроса (теоремы, примера и т.п.), и только в одном смысле. Таким образом, в границах выделенного структурного элемента можно уверенно предполагать, что формульный контекст является постоянным, и символическая конструкция, обозначающая переменную, в рамках одного структурного элемента имеет один смысл.

1.2. Поиск

В алгоритме поиска используются два типа сущностей: набор словосочетаний, определенных пользователем в поисковом запросе как названия некоторых величин, взаимосвязь между которыми в виде математической формулы требуется найти, и математические конструкции в размеченном документе. Для работы с математическими конструкциями используются элементы математической разметки. Для работы со словосочетаниями, составляющими поисковый запрос, используются методы поиска в сплошных текстах.

Реализация поиска может быть выполнена двумя способами. В первом варианте поиск начинается с выполнения полнотекстового запроса, затем следует фильтрация результатов и использование разметки математического текста. Во втором варианте на основе математической разметки проводится отбор документов, которые могут содержать искомые формулы, и далее выполняется фильтрация документов по полнотекстовым запросам. Выбор того или иного варианта зависит от типа коллекции, по которой производится поиск. Если поиск производится по разнородной коллекции, например, Интернету в целом, то второй вариант может оказаться предпочтительнее, так как количество документов, содержащих математические формулы, значительно меньше общего числа документов коллекции. Если поиск производится по специализированным математическим коллекциям, то более целесообразно начать с выполнения полнотекстового поиска.

Полнотекстовый поиск словосочетаний поискового запроса обладает некоторыми особенностями. Словосочетания запроса, именующие естественнонаучные величины, являются именными группами с достаточно сложной внутренней структурой (например, многословный термин «допустимый прогиб квадратной деревянной балки при сосредоточенной нагрузке»). Обнаружив в тексте словосочетание из запроса и математическую конструкцию в непосредственной близости, можно предположить, что локализован фрагмент документа с текстовым определением символьной конструкции.

С другой стороны, культура написания технических текстов допускает, определив предмет обсуждения, именовать относящиеся к нему величины без упоминания самого объекта рассмотрения. Так, если статья называется «О квадратных деревянных балках под сосредоточенной нагрузкой» определение интересующей нас переменной может быть дано кратко, например как «допустимый прогиб» или «допустимый прогиб балки». Таким образом, полнотекстовый поиск должен отмечать как релевантные все вхождения не только исходного определения, но и всех возможных усечений именной группы, вплоть до единственного главного слова именной группы. С другой стороны, вхождение большого количества и, даже всех, слов исходного словосочетания в текст, но не в соответствующей одному из усечений именной группы форме, не должен рассматриваться как релевантный.

Определив позиции релевантных вхождений словосочетаний запроса в анализируемый текст, далее исследуется гипотеза о наличии определения символьного обозначения некоторой величины. Для этого необходимо проанализировать, с каким из вхождений фрагментов математической разметки, соответствующих отдельной переменной,

соотносится данный текст. Это довольно сложный вопрос, требующий отдельного рассмотрения. Стилистика естественнонаучных текстов допускает достаточно произвольное произвольные формы определения символьных обозначений, наряду с общеупотребительными конструкциями типа *"... , где x - допустимый прогиб..."*, или *"... пусть x - допустимый прогиб..."*, определения могут иметь более сложную структуру, например, иметь табличную структуру, в которой кроме обозначений содержится дополнительная информация. Несколько подряд идущих определений ставят вопрос об ассоциативности: с какой ближайшей символьной конструкцией должно быть ассоциировано вхождение словосочетания запроса. В настоящее время выявлено несколько эвристик, однако этот вопрос нуждается в дальнейших исследованиях.

После установления связи между словосочетаниями запроса и символьными конструкциями производится проверка, входят ли данные символьные конструкции в одну математическую формулу. Эта задача решается на основе построенной разметки формулы, в которой с каждым вхождением переменной в тексте ассоциируются все ее вхождения в математические формулы, имеющиеся в тексте. Если в анализируемом тексте имеются одна или несколько математических формул, в которой использованы все символьные определения. соответствующие словосочетаниям запроса, то найденный текст полностью релевантен запросу.

Случаи частичной релевантности требуют дополнительного рассмотрения. Возможна частичная релевантность нескольких типов: установлено наличие формул, связывающих не полные словосочетания запроса, а некоторые их усечения; присутствует формула, связывающая часть словосочетаний запроса. Каждый из этих случаев требует своего метода сортировки по релевантности и метода их взаимного ранжирования. Решение этих вопросов требует проведения вычислительных экспериментов на реальных коллекциях.

2. Прототип системы семантического поиска математических выражений

Для проверки базовых концепций поиска математических формул была реализован прототип системы семантического поиска математических выражений. разметку, весьма близкую к TeX - разметке. параметров, а также ссылка на страницу Википедии в Интернете. Прототип системы специализирован для поиска математических формул в статьях русской Википедии. Выбор Википедии для экспериментов связан, с одной стороны, с тем, что Википедия является одной из крупнейших коллекций научных текстов из различных областей знаний, с другой стороны, математические выражения в Википедии имеют унифицированную форму представления. В текущей реализации запрос поддерживает ввод до пяти наименований текстовых параметров формулы (рис. 2). Результатом поиска является список страниц Википедии, на которых отображается формула, фрагменты содержимого страницы с определением заданных параметров, а также ссылка на страницу Википедии в Интернете.

2.1. Архитектура системы

На рис. 3 представлена архитектура разработанного прототипа системы семантического поиска математических выражений.

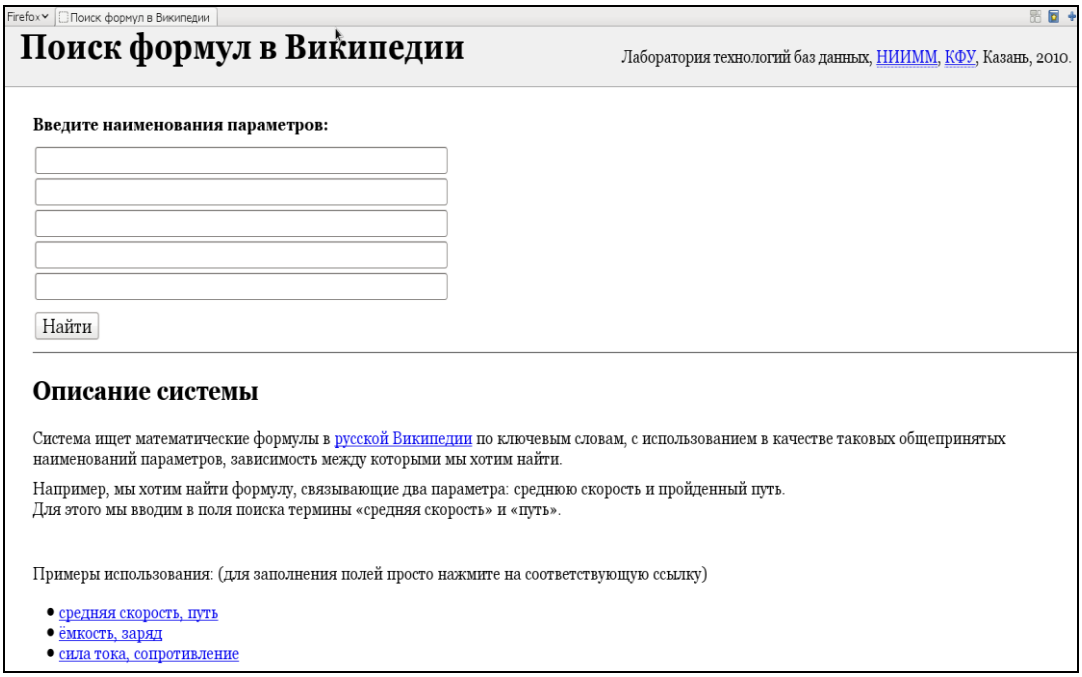

Рисунок 2 – Web-страница ввода параметров семантического поиска формул

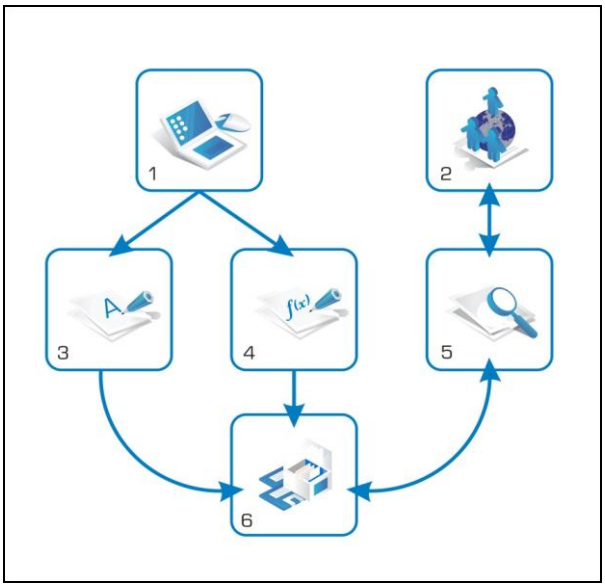

Рисунок 3 – Архитектура прототипа системы семантического поиска математических выражений

Прототип включает следующие взаимосвязанные подсистемы (рис. 3):

- подсистема загрузки и анализа данных Википедии;
- подсистема взаимодействия с пользователем;
- подсистема полнотекстового индексирования;
- подсистема индексирования математических формул и переменных;
- подсистема поиска и ранжирования;
- подсистема хранения данных.

2.1.1. *Подсистема загрузки и анализа данных*

Технически возможно несколько вариантов загрузки исходной информации (постраничное копирование статей Википедии в реальном масштабе времени или формирование архива статей Википедии в xml –формате). В настоящий момент реализован второй вариант как наиболее предпочтительный на этапе разработки, т.е. первоначально производится импорт данных русской Википедии из предварительно полученного архива. Во время импорта осуществляется анализ и обработка загружаемых данных. Единицей анализируемой и загружаемой информации является html-страница. Страница разбирается на структурные части, выделяется заголовок и контент. Процедура разбора проводится с помощью стандартного Java SAX-парсера. Затем контент страницы анализируется на наличие математических формул. Анализ проводится стандартными средствами Java и будет описан далее. Страницы, содержащие формулы, сохраняются в базе данных для дальнейшего использования, остальные отбрасываются. Завершается подготовка системы к работе индексированием сохраненной информации для обеспечения возможности поиска.

2.1.2. *Подсистемы индексирования*

Для обеспечения возможности высокоскоростного поиска ключевых фраз и математических формул необходимо осуществлять их индексирование. В системе производится индексирование входных документов как текстовых данных, а также дополнительное индексирование математических формул.

Для решения задачи поиска релевантных документов по набору ключевых слов (полнотекстового поиска) используется библиотека Apache Lucene [Lucene]. Подсистема хранения Lucene, как и в большинстве современных поисковых систем, организована в виде так называемого обратного или инвертированного индекса. Обратный индекс представляет собой список слов, документированных в алфавитном порядке с указанием позиции и других параметров вхождения слова, т. е. организован аналогично предметному указателю в конце книги – каждому слову соответствует список документов, в которых он встречается. Такая структура позволяет практически за константное время извлекать список документов, в которых встречается определенное слово. Кроме собственно номеров документов для каждого слова сохраняется ряд атрибутов, таких как все позиции, в которых встречается данное слово в данном документе, сдвиги относительно предыдущего слова, ряд других встроенных атрибутов, а также дополнительные атрибуты. Индекс Lucene фактически представляет собой документно-ориентированную базу данных с объектами и полями этих объектов. В отличие от реляционных баз объекты в ней не обязаны подчиняться какой-либо схеме. Каждый документ может иметь произвольное число произвольных полей, независимо от того, какие поля имеют другие документы.

Для индексирования математических фрагментов выполняются следующие операции: обнаружение в тексте; классификация (формула, переменная, другие); построение позициональных индексов формул и переменных; построение индексов соответствия переменных формулам.

Для записи математической формулы в Википедии используется формат записи формул LaTeX. В настоящее время для обработки математических текстов доступны ряд открытых библиотек (Jeuclid [http://jeuclid.sourceforge.net] - поддерживает формат MathML; TeXlipse [http://texlipse.sourceforge.net] – формат LaTeX; SnuggleTeX [http://www2.ph.ed.ac.uk/snuggletex] поддерживает преобразование LaTeX в MathML). Перечисленные инструменты по ряду причин не используются в настоящей реализации (требуется модификация исходного кода инструментов, существует зависимость результатов разбора формул от целевой функции и др.), однако в дальнейшем планируется их использование для более полного разбора формул. Прототип системы поиска использует оригинальный алгоритм

обнаружения и классификации математических фрагментов. Фрагментом формулы считается любой текст между специализированными тегами разметки *$$* (рис.4).

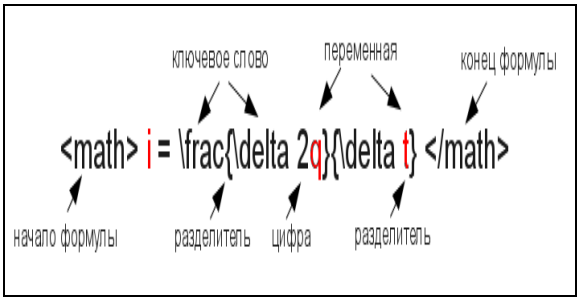

Рисунок 4 – Структура математической формулы

Полученный фрагмент очищается от служебных символов языка разметки Википедии, лишних пробельных символов. Далее фрагмент проверяется на соответствие ряду критериев (длина, количество переменных, наличие операторов отношений и операций). Если фрагмент удовлетворяет основным критериям, то он считается формулой (переменной или другим, например, таблицей). Позиции формул и переменных в тексте запоминаются в соответствующих индексах.

При построении индексов соответствия формул и переменных важным является наличие уникальных переменных в формуле, поэтому анализ формулы значительно упрощается (в отличие от полного грамматического разбора). В качестве инструмента анализа используется язык регулярных выражений. Сначала формула разбивается на фрагменты, разделителями считаются различные символы скобок, символы арифметических и логических операций, знаки пунктуации, пробельные символы и т. п. Полученные фрагменты анализируются на принадлежность к специальным группам (ключевые слова (начинаются с символа «\»), нижние индексы (начинаются с символа «_»), числа и т. п.). Если фрагмент на этом этапе не классифицирован, то с большей долей вероятности его можно считать переменной. Выявленные ранее переменные в тексте и выявленные переменные в формулах проверяются на соответствие, затем строится индекс вхождения переменных в формулы.

2.1.3. *Подсистема поиска и ранжирования*

Процесс решения задачи семантического поиска математических выражений выполняется в несколько этапов. На первом этапе производится полнотекстовый поиск всех вхождений ключевых словосочетаний в тексты. Для каждого вхождения определяется, есть ли в некоторой окрестности ключевой фразы (не более 50 символов) переменная. По найденным переменным определяется соответствующая формула. Для каждой формулы строится группа текстовых фрагментов, включающих ключевые словосочетания и переменные.

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

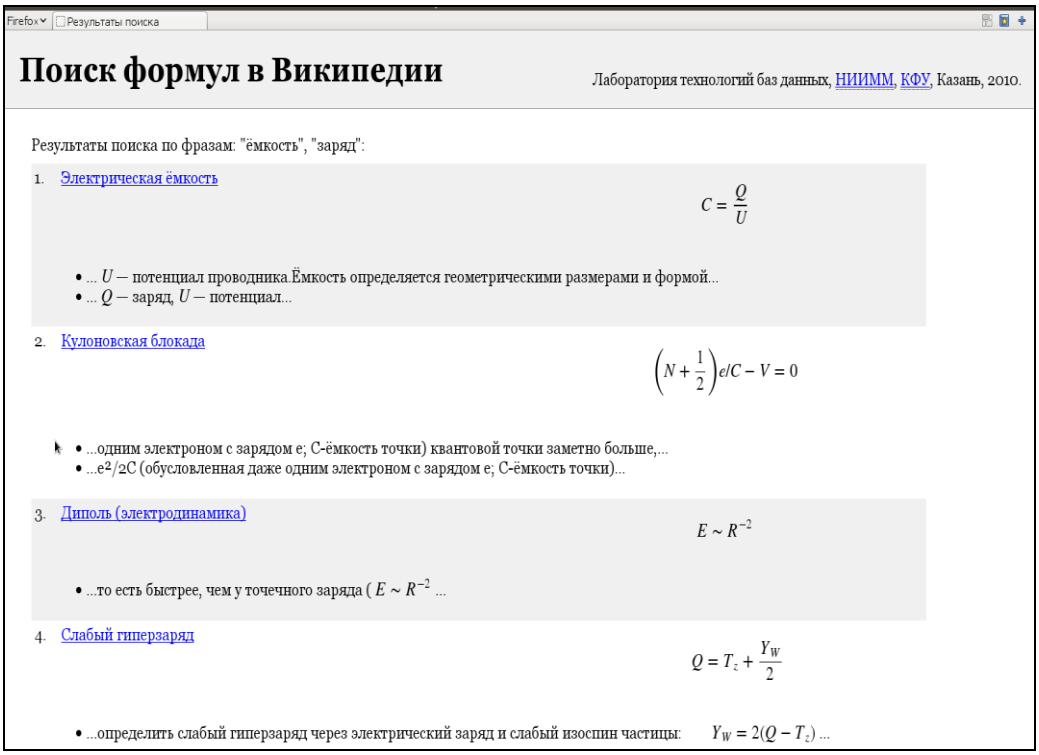

Рисунок 5 – Результаты математического поиска по параметрам “емкость”, “заряд”

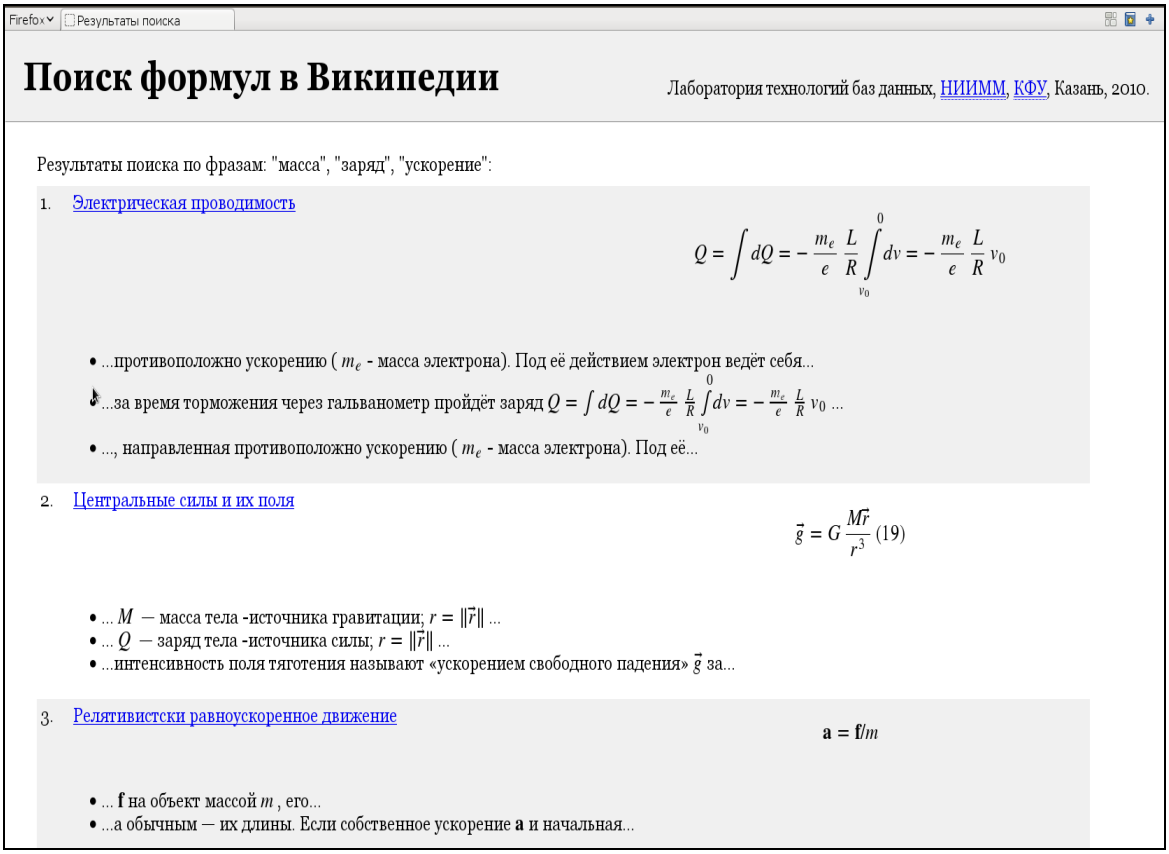

Рисунок 6 – Результаты математического поиска по параметрам “масса”, “заряд”, “ускорение”

фрагментов с определениями переменных от позиции соответствующей формулы. Предполагается, таким образом, что формула и определения, входящих в них переменных, должны быть достаточно близки друг к другу. В результате для каждого документа получаем оптимальную группу текстовых фрагментов (потенциальных определений) и относящуюся к ним формулу. Результаты для всех документов сортируются по критерию близости и дополнительным критериям релевантности.

3. Программная реализация

Пользовательский интерфейс представляет собой Web-приложение, доступ к которому осуществляется через любой современный браузер с поддержкой JavaScript. Интерфейс состоит из двух Web-страниц: страница ввода запросов пользователей и страницы представления результатов поиска.

На странице ввода запроса (рис. 2) пользователь может указать в текстовых полях до пяти названий параметров, которые должны присутствовать в искомой формуле, и после этого инициировать поиск. Дополнительно на странице можно выбрать предопределенные запросы («*Примеры использования*») и информацию программе («*Подробнее о программе*»).

Результаты поиска отображаются на странице результатов (рис. 5, 6), где выводится запрос и ранжированный по релевантности список ссылок на страницы Википедии.

Для каждой ссылки приводятся наиболее вероятная формула и фрагменты, содержащие параметры запроса. Так, например, на рис. 6 приведены результаты поиска для параметров «*масса*», «*заряд*», «*ускорение*». Результатом поиска являются формулы, связывающие физические величины запроса.

В качестве основной программной платформы реализации выбрана кроссплатформенная технологии Java 1.6. Все используемые в системе сторонние продукты и библиотеки являются свободно распространяемыми открытыми решениями. Сервер приложений организован на основе встроенного Apache Tomcat 7.0 [http://tomcat.apache.org], а в качестве хранилища данных выбрана СУБД PostgreSQL 8.4 [http://www.postgresql.org], но также возможно использование других СУБД при использовании соответствующих драйверов. Объектная модель системы связывается с таблицами базы данных посредством ORM-технологии Apache OpenJPA [http://openjpa.apache.org].

Выполнение основных стандартных операций полнотекстового поиска по заданным ключевым

фразам производится с помощью библиотеки Apache Lucene [http://lucene.apache.org].

Для отображения найденных в документах формул возможно использование как библиотеки JlaTeXMath [http://forge.scilab.org/index.php/p/jlatexmath] преобразования текстовой записи формулы в изображение png-формата, так и библиотеки MathJax [www.mathjax.org] динамического отображения формул в Web-контенте.

Заключение

В статье рассмотрен подход к семантическому поиску математических выражений, позволяющий выполнять запросы на поиск математических формул по текстовым наименованиям переменных, входящих в формулы.

В заключение обсудим ряд важных проблем будущих исследований. Одной из сложных задач является разработка методов разрешения многозначности, связанной с множественными вхождениями в текст определений переменных. На первый взгляд представляется, что эта проблема может существенно снизить релевантность поиска, однако можно привести некоторые возражения. Аналогичная проблема изучалась при рассмотрении задачи организации поиска в базах данных на основе запросов на естественном языке [Бирюльцев и др., 2006], [Бирюльцев и др., 2007]. В этой задаче рассматривалась многозначность при ассоциации логических (пользовательских) наименований полей БД с физическим полями в конкретном экземпляре БД. Однако, было показано, что логика самой БД в виде группировки полей в таблицы и миграции ключей между таблицами отсеивает нерелевантные сочетания [Бирюльцев и др., 2007а]. Можно предположить, что при наличии множества определений одной и той же символьной конструкции в тексте, логика предметной области устранит большинство нерелевантных сочетаний этих неоднозначно трактуемых символьных обозначений. Достаточно маловероятно, что в одном и том же тексте автор так переопределил используемые им символьные обозначения, что они одновременно появятся в формулах, имеющих различный смысл. Таким образом, проблему множественности определений символьных обозначений, весьма сложную и требующую привлечения лингвистического и семантического анализа, возможно, не придется решать вовсе, а разрешение неоднозначности будет, в подавляющем числе случаев, выполняться автоматически на этапе поиска. Окончательная проверка данного предположения может быть проведена на основе численного эксперимента с коллекциями математических текстов.

Рассмотренный в статье прототип системы, несмотря на некоторые функциональные упрощения, показал принципиальную работоспособность предложенного подхода. Реализованный в системе алгоритм поиска математических формул по введенным ключевым фразам (названиям параметров формулы), в процессе эксплуатации показал достаточную релевантность в сочетании с высокой скоростью поиска. Анализ результатов поиска показал, что выдаваемые результаты практически всегда имеют непосредственное отношение к задаваемому запросу, и на первой странице поиска находится формула, отвечающая запросам пользователя.

Однако можно указать и некоторые наиболее существенные проблемы реализации, которые будут исследованы в дальнейшем. В первую очередь необходимо дополнить методы ранжирования фрагментов, содержащих переменную и поисковый запрос, частичным синтаксическим анализом. Это позволит отфильтровывать конструкции, не имеющие синтаксических признаков определения, а также производить более адекватное аннотирование найденных определений.

В ближайшей перспективе планируется провести экспериментальное оценивание разработанного подхода, в том числе, оценку эффективности алгоритмов поиска и ранжирования в терминах стандартных метрик точности, полноты и nDCG [Jarvelin et al., 2000].

Также ведется работа по интеграции предложенной модели поиска по формулам с поиском по структурным элементам и терминологии [Бирюльцев и др., 2010] для более точного задания и разрешения поисковых запросов математической предметной области.

Благодарности

Работа выполнена при финансовой поддержке РФФИ, грант № 11-07-00507а.

Библиографический список

[Бирюльцев и др., 2006] Бирюльцев Е. В., Гусенков А. М., Галимов М. Р. Особенности лексико-семантической структуры наименований артефактов реляционных баз данных // Тр. Казан. школы по компьютерной и когнитивной лингвистике TEL'2005. – Казань: Изд-во Казан. ун-та, 2006. – Вып. 9. – С. 4-12.

[Бирюльцев и др., 2007] Бирюльцев Е.В., Гусенков А.М., Елизаров А.М. О доступе к электронным коллекциям в виде реляционных баз данных на основе онтологий //Труды 9-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007, Переславль-Залесский, Россия, 15–18 октября 2007 г. – Переславль-Залесский : Изд-во «Университет города Переславля», 2007. – С. 211– 216.

[Бирюльцев и др., 2007а] Бирюльцев Е.В., Гусенков А.М. Интеграция реляционных баз данных на основе онтологий. // Ученые записки Казанского государственного университета. Серия Физико-математические науки. – 2007. – Том 149. – Книга 2. – Казань: Казан. ун-т, 2007. – С. 13 –25.

[Бирюльцев и др., 2010] 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 XII Всероссийской

научной конференции RCDL'2010. – Казань: Казан. ун-т. – 2010. – С. 296–300.

[Google] Google Scholar. URL: http://scholar.google.com

[Jarvelin et al., 2000] Jarvelin, K., Kekalainen, J. IR evaluation methods for retrieving highly relevant documents // Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. – ACM. – 2000. – P. 41–48.

[**Kohlhase M., 2005**] Kohlhase M. sTeX: Semantic Markup in TeX/LaTeX. – 2005. – Режим доступа: https://svn.kwarc.info/repos/stex/trunk/sty/stex.pdf

[**Lange C., 2009**] Lange C. Krextor – An Extensible XML -> RDF Extraction Framework // CEUR Workshop Proceedings. – 2009 – V. 449.

[LatexSearch] The Springer LaTeX Search. URL: http://latexsearch.com

[Lucene] Apache Lucene. URL: http://lucene.apache.org

[MathWebSearch] MathWebSearch. URL: http://search.mathweb.org/index.xhtml

[Microsoft] Microsoft Academic Search. URL: http://academic.research.microsoft.com

[**Miller B., 2007**] Miller B. LaTeXML: A LaTeX to XML converter. – 2007. – Режим доступа: http://dlmf.nist.gov/LaTeXML

[Misutka et al., 2008] Misutka, J., Galambos, L. Extending Full Text Search Engine for Mathematical Content // Proceedings of DML. – 2008. – P. 55–67.

[RDFa, 2008] RDFa in XHTML: Syntax and processing. Recommendation W3C. – 2008.

[Solovyev et al., 2010] Solovyev V., Zhiltsov N. Logical structure analysis of scientific publications in mathematics // Proceedings of the International Conference on Web Intelligence, Mining and Semantics (WIMS). – ACM. – 2011. - P. 21:1-21:9

[Uniquation] (uni)quation. URL: http://uniquation.ru

A NOVEL APPROACH TO SEMANTIC SEARCH OF MATHEMATICAL EXPRESSIONS IN THE SCIENTIFIC DOCUMENTS

Birialtsev E.V. [*], Galimov M.R. [*], Zhiltsov N.G. [*], Nevzorova O.A. [**]

** Kazan (Volga Region)Federal University, Kazan, Russia*

IgenBir@yandex.ru

glmvmrt@gmail.com

nikita.zhiltsov@gmail.com

*** Research Institute of Applied Semiotics of the Academy of Sciences of Tatarstan Republic, Kazan, Russia*

onevzoro@gmail.com

We propose a novel approach to semantic search of mathematical expressions. It enables users to query scientific documents (e.g. scholarly papers or Wikipedia web pages) using textual representation of variables involved in the relevant formulas. We also present a prototype implementing semantic search of mathematical expressions. As a conlusion, we discuss a few current issues of the present implementation and future work.

INTRODUCTION

Currently, there are a lot of approaches to search of mathematical formulas that slightly differ from each other. Most of them require either proficiency in formalizing a search query in terms of the specific representation languages, such as LaTeX, or dealing with cumbersome UI query editors. However, none of them attempt to elicit semantic relationships between mathematical symbols and their corresponding textual definitions. Thus, our approach aims to bring the gap between conventional keyword search and mathematical formula search by leveraging the semantics related to mathematical variables involved in a given mathematical expression.

The main task while searching mathematical formulas is parsing initial expressions in the representation formats and converting them into semantically enriched models. As a matter of fact, MathML and OpenMath are two standard languages that provide capabilities to capture the underlying semantics of mathematical expressions including the formula structure, involved variables and properties of operations between them. In practice, particular tools are used to convert from LaTeX based representation into MathML.

Given a full text index and MathML represented formulas, our first specific task is to associate a term or a group of terms from the index with an expression variable symbol. Finally, we should provide possibility for a user to query using indexed terms and retrieve relevant formulas that contain associated variables.

MAIN PART

We use Content and Presentation layers of MathML as internal representation formats for parsed formulas. Besides, our full text indexing machinery is based on the Apache Lucene library.

A set of our core algorithms includes indexing the document text contents relying on the document logical structure, parsing LaTeX expressions, detecting symbol definitions and connecting them with stored formulas. At the moment, detecting local contexts of variables performs runtime while querying. During the query resolution process, our method seeks over all the text definitions as a fixed-sized token window around occurrences of a given variable.

Our search prototype provides opportunities to query the search index built upon a Wikipedia page dump using a concise user interface and view search snippets enriched with the local context information related to a given relevant formula.

CONCLUSION

Among main issues of our approach, there are low disambiguation of particular text definition parts and insufficient robustness of ranking mechanisms to marginally useful keywords.

We are going to study efficiency of our approach on the real-world use cases and conduct experiments on evaluating its performance using several standard search quality metrics.

We also have some work in progress on incorporating the document structural analysis mechanisms into our formula search prototype.