УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО УНИВЕРСИТЕТА
Том 156, кн. 4 Физико-математические науки *2014*

УДК 004.82+004.89+004.912

НЕКОТОРЫЕ ПОДХОДЫ К РАЗМЕТКЕ ЕСТЕСТВЕННОНАУЧНЫХ ТЕКСТОВ, СОДЕРЖАЩИХ МАТЕМАТИЧЕСКИЕ ВЫРАЖЕНИЯ

Е.В. Биряльцев, А.М. Гусенков, О.Н. Жибрик

Аннотация

В работе анализируется подход к семантическому поиску математических выражений, позволяющий выполнять запросы на поиск математических формул по текстовым наименованиям переменных, входящих в формулы. Предлагается метод установления связей между текстовыми определениями переменных, их обозначениями и формулами, в состав которых эти переменные входят. Описываются две реализации систем поиска: поиск формул в статьях интернет-энциклопедии Википедия и разметка корпусов математических текстов для поиска по онтологии. Обсуждаются результаты оценочных экспериментов с точки зрения релевантности поиска и полноты связывания, а также способы решения основных проблем предложенного подхода.

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

Введение

Поиск по математическим документам – актуальная и быстроразвивающаяся область исследований. С одной стороны, хорошо известные системы, такие как Google Scholar [1], Microsoft Academic Search [2], реализуют полнотекстовый поиск по ключевым словам в научных коллекциях и интернет-ресурсах, который весьма удобен для конечного пользователя. С другой стороны, ряд специализированных систем поиска по математическим формулам предлагает средства для формулирования запроса в синтаксисе языка разметки LaTeX (например, Springer LaTeXSearch [3] индексирует базу статей издательства Springer; (uni)quation [4] – научные тематические сайты, форумы и Wikipedia) или MathML, используя соответствующие графические интерфейсы [5].

Подход, представленный в настоящей статье, направлен на интеграцию функциональных возможностей полнотекстового поиска и поиска по математическим формулам, при котором конечному пользователю предлагается формулировать запрос на поиск математической формулы в форме ключевых слов. Наиболее близким к предлагаемому 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 [6].

Новизна предлагаемого подхода состоит в том, что в качестве поискового объекта рассматривается сложный нелинейный нетекстовый объект, включающий собственно математическое выражение формулы в нотации одного из соответствующих языков представления и набор определений символьных обозначений, участвующих в математическом выражении, которые извлекаются из всего анализируемого текста.

Данная постановка позволяет рассматривать предлагаемый подход как новый тип запросов к коллекциям математических документов. Действительно, в отличие от известной задачи поиска математической формулы по её фрагменту, в формулировке запроса должны использоваться не математические конструкции, а словесные наименования переменных, входящих в искомую формулу.

В настоящей статье рассматриваются два способа разметки научных текстов, содержащих формулы, для выполнения запросов на естественном языке. Оба способа основаны на выделении в тексте формулы и входящих в неё переменных в формате LaTeX.

Первый подход связан с организацией поиска в статьях русскоязычной части популярной универсальной интернет-энциклопедии Википедия, заключающейся в индексации документов по ключевым словам и связывании с ними формульных выражений.

Второй подход был использован при построении онтологии на основе связной коллекции математических документов, включающей в себя, кроме самих компонентов онтологии, связанные с ними переменные и формульные выражения [7, 8]. Поисковый запрос пользователя на естественном языке переводится в термины онтологии, по которым затем формируется запрос на языке запросов к RDF-документам SPARQL [9].

В обоих случаях результатом поиска, в отличие от полнотекстовых запросов, являются не просто фрагменты документа из коллекции, содержащие слова строки запроса в определенном количестве и находящиеся в достаточной близости, а наборы фрагментов двух видов: фрагментов, содержащих нетекстовый объект (формулу), и фрагментов, связывающих переменные формулы и их текстовые определения вне зависимости от их местонахождения в тексте. Для выполнения поискового запроса в анализируемом тексте выделяются переменные по их текстовому описанию и формулы, в которых данные переменные представлены наиболее полно. Таким образом, указанный тип запроса не сводится ни к поиску формул по их фрагментам, ни к полнотекстовому поиску, что потребовало применения оригинальных подходов, особенно в части методов разметки и индексирования коллекций.

Подробное описание предлагаемого подхода было изложено в [7, 8, 10, 11, 12–15], там же было проведено его сопоставление со смежными работами. В настоящей работе впервые представлены экспериментальные данные апробации такого подхода на корпусе математических текстов.

1. Подход к семантическому поиску математических выражений в Википедии

В рассматриваемой постановке задачи в естественнонаучных текстах выделяются следующие виды сущностей: естественнонаучные термины, символьные условные обозначения терминов, математические фрагменты (формулы). Все перечисленные сущности составляют контекст формулы.

Таким образом, для решения поставленной задачи необходимо вычленять в тексте 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 Википедии 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 сплошном тексте, а формула, которая связывает данные переменные, представляет собой нетекстовый объект.

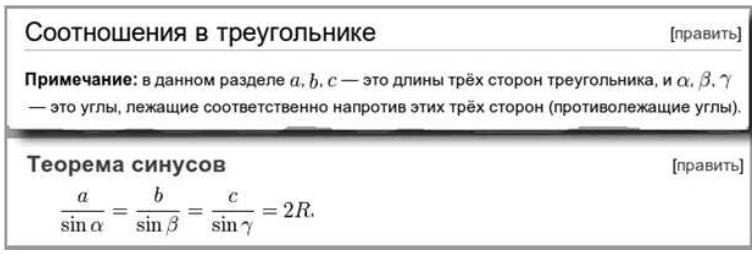


Рис. 1. Пример формульного контекста (фрагмент статьи Википедии)

Исходя из методов реализации задачи поиска в базах данных по запросу на естественном языке [10, 16, 17], нами предложен подход к решению данной задачи [7, 11], использующий дополнительную разметку и индексирование математических текстов. Дополнительная разметка и индексирование включает в себя конструкции, указывающие связи математических формул и обозначений символьных переменных, участвующих в этих формулах. Данная разметка, или соответствующий ей индекс, учитывается на этапе поиска следующим образом. Поиск заданных в поисковом запросе текстовых наименований переменных производится только в окрестностях символьных переменных, находящихся в тексте. Предполагается, что появление текстового наименования переменной в окрестностях символьного представления некоторой переменной указывает на некоторую семантическую связь между ними, в частности может являться определением обозначения этой символьной переменной через её текстовое наименование. Таким образом, мы можем найти символьное обозначение переменной через её заданное пользователем в поисковом запросе текстовое наименование. Далее проверяются связи выделенных символьных переменных через одновременное вхождение в математические формулы. Математические формулы, в которые входят все найденные переменные, предоставляются пользователю совместно с фрагментами текста, содержащими текстовое и символьное наименование переменных как результат поискового запроса.

1.1. Система семантического поиска математических выражений. Для проверки базовых концепций предложенного метода была реализована система семантического поиска математических выражений, специализированая для поиска математических формул в статьях русской Википедии. Выбор Википедии был связан, с одной стороны, с тем, что Википедия является одной из крупнейших коллекций научных текстов из различных областей знаний, с другой – математические выражения в Википедии имеют унифицированную форму представления.

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.

Система включает следующие взаимосвязанные подсистемы:

- 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.

1.1.1. Подсистема загрузки и анализа данных. Технически возможны несколько вариантов загрузки исходной информации (постраничное копирование

статей Википедии в реальном масштабе времени или формирование архива статей Википедии в xml-формате). В системе реализован второй вариант, то есть 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. Страницы, содержащие формулы, сохраняются в базе данных для дальнейшего использования, остальные отбрасываются. Завершается подготовка системы к работе индексированием сохраненной информации для обеспечения возможности поиска.

1.1.2. 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 [18]. Подсистема хранения Lucene, как и в большинстве современных поисковых систем, организована в виде так называемого обратного или инвертированного индекса, аналогичного предметному указателю в конце книги, – каждому слову соответствует список документов, в которых он встречается. Такая структура позволяет практически за постоянное 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, а также дополнительные атрибуты. Индекс Lucene фактически представляет собой документно-ориентированную базу данных с объектами и полями этих объектов.

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 доступен ряд открытых библиотек. Например, Jeuclid [http://jeuclid.sourceforge.net] поддерживает формат MathML, TeXlipse [http://texlipse.sourceforge.net] работает с форматом LaTeX, SnuggleTeX [http://www2.ph.ed.ac.uk/snuggletex] позволяет преобразовывать формат LaTeX в MathML. Перечисленные инструменты по ряду причин (требуется модификация исходного кода инструментов, существует зависимость результатов разбора формул от целевой функции и др.) не используются в рассматриваемой реализации, однако в дальнейшем возможно их использование для более полного разбора формул.

Подсистема поиска использует оригинальный алгоритм обнаружения и классификации математических фрагментов. Фрагментом формулы считается любой текст между специализированными тегами разметки *$$* (рис. 2).

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 математическим выражением (формулой, переменной или другим типом,

Рис. 2. Структура математической формулы

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.

1.1.3. 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, существует 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 фрагментов с определениями переменных от позиции соответствующей формулы. Предполагается, таким образом, что формула и определения входящих в неё переменных должны быть достаточно близки друг к другу. Для определения достаточной близости вводится понятие максимально допустимого расстояния (МДР), способы определения которого представляют собой отдельную задачу. На начальном этапе МДР может определяться эмпирическим путём с последующим уточнением на основе статистических данных. В результате для каждого документа получаем оптимальную группу текстовых фрагментов (потенциальных определений) и относящуюся к ним формулу. Полученные данные для всех документов сортируются по критерию близости и дополнительным критериям релевантности.

1.1.4. Пользовательский интерфейс. Пользовательский интерфейс представляет собой Web-приложение, доступ к которому осуществляется через любой современный браузер с поддержкой JavaScript. Интерфейс состоит из двух Web-

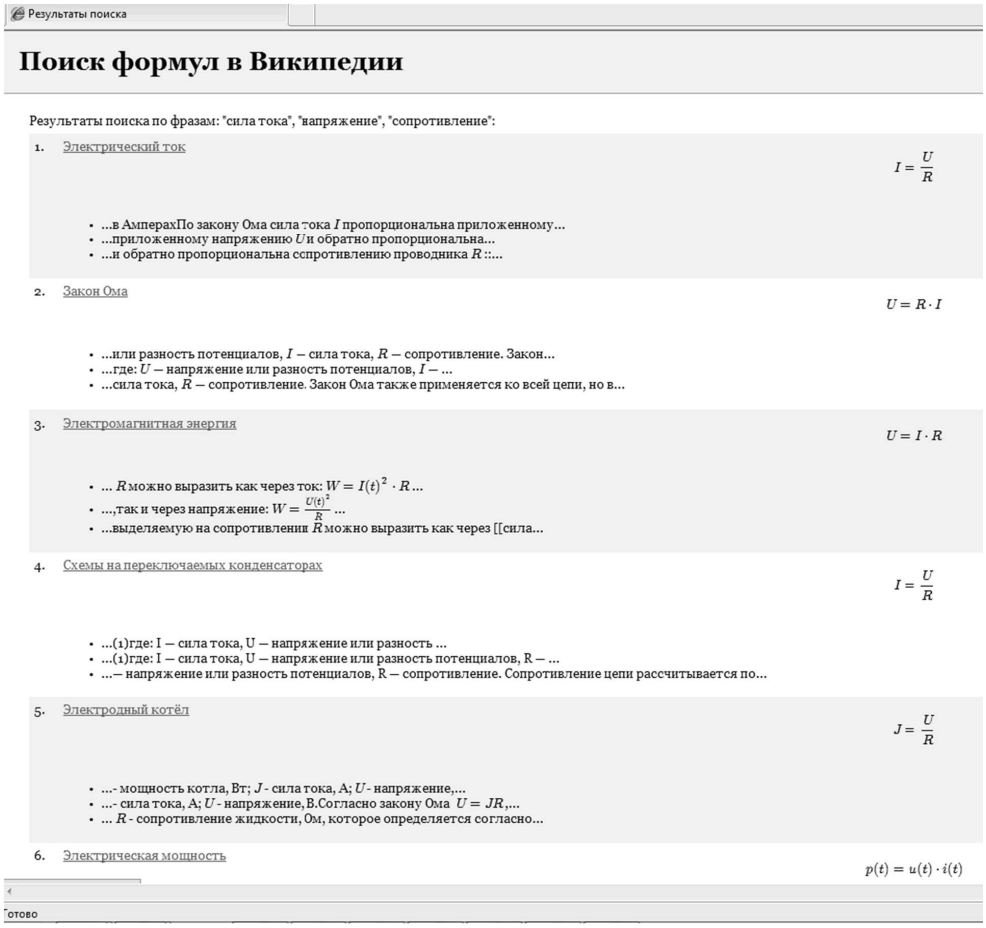


Рис. 3. Закон Ома

страниц: страница ввода запросов пользователей и страницы представления результатов поиска.

На странице ввода запроса пользователь может указывать в текстовых полях одно или более названий параметров, которые должны присутствовать в искомой формуле, и после этого инициировать поиск.

Результаты поиска отображаются на странице результатов, где выводится запрос, а также ранжированный по релевантности список ссылок на страницы Википедии.

1.2. Экспериментальная оценка релевантности. Был проведен ряд численных экспериментов для выявления основных проблем с полнотой и релевантностью поиска.

В процессе экспериментов анализировались результаты поиска законов путем задания наименования параметров, входящих в искомый закон. Задавались как хорошо известные законы, например закон Ома (рис. 3), так и законы из специальных областей знаний (рис. 4, закон Дарси, подземная гидромеханика). Искомые формулы были найдены во всех 10 исследованных случаях. Релевантность ответов также достаточно хорошая, в 8 случаях из 10 искомый закон располагался на 1-й сроке списка выдачи.

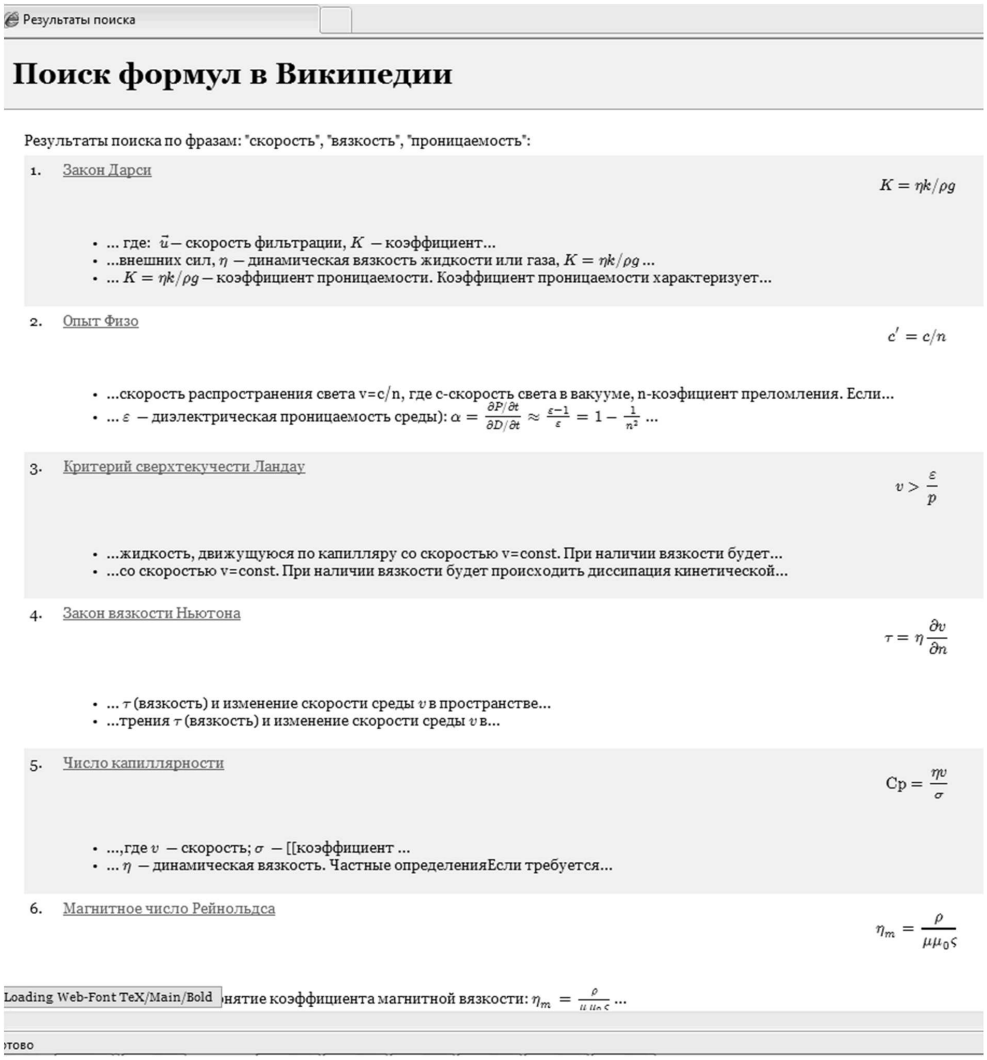


Рис. 4. Закон Дарси

1.3. Обсуждение результатов экспериментов. Реализованный в системе алгоритм поиска математических формул по введенным ключевым фразам (названиям параметров формулы) 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.

Вместе с тем были выявлены проблемы, для решения которых требуется улучшение механизма поиска. Наглядным примером этих проблем является низкая релевантность ответов по запросам, связанным с законами Ньютона. Анализ статей Википедии, которые выдавались как более релевантные по данным запросам, показал следующее. Большинство выдаваемых статей касалось механики сплошных сред, небесной механики, теории относительности и других разделов физики и механики, в которых рассматривались различные частные случаи механического движения. Ключевые слова «масса» и «сила» употреблялись в них чаще, чем

в собственно искомой статье. Вместе с тем эти ключевые слова употреблялись преимущественно в составе именных групп «приливные силы», «центробежные силы», «релятивистская масса», «масса покоя» и т. д.

Таким образом, приведенный пример показывает, что для дальнейшего улучшения релевантности необходимо модифицировать механизм поиска, а не механизм ранжирования. Наиболее очевидным путем представляется синтаксический анализ текста с целью выделения именных групп и дальнейшего анализа отношения выделенных именных групп и заданных пользователем наименований параметров, также являющихся именными группами. Вариантов отношений, исключая тривиальный случай полного несовпадения входящих в именные группы слов, насчитывается четыре:

1) именные группы полностью совпадают;
2) именная группа параметра входит в именную группу текста;
3) именная группа текста входит в именную группу параметра;
4) именные группы имеют общее подмножество слов.

Первый вариант является условием наибольшей релевантности. Во втором варианте речь в тексте идёт о более частном случае по сравнению с заданным пользователем. В третьем варианте пользователь, напротив, интересуется более частным случаем, чем тот, о котором говорится в тексте. В четвёртом варианте речь идёт о различных частных случаях.

Таким образом, на этапах индексирования и поиска необходимо учитывать взаимные отношения именных групп, находящихся в тексте в составе возможных определений символьных переменных, и именных групп, заданных пользователем в строках запроса.

Второй выявленный при экспериментах аспект касается поиска формулы по наименованию самой формулы. Как показал тот же случай с поиском законов Ньютона, явное указание наименования закона, если оно известно пользователю, позволяет найти требуемую формулу проще и с более высокой релевантностью. Вместе с тем реализованный алгоритм не позволяет указать явно, что задаваемая поисковая строка является наименованием закона, а не входящего в него параметра.

Реализация такой возможности не требует принципиальной переработки поисковых механизмов. Достаточно будет указать, что некоторый элемент поискового запроса должен находиться в окрестности математического выражения.

При реализации поиска по наименованию закона также будет полезно для повышения информативности представляемого результата указывать обозначения входящих в найденную формулу переменных. Это является новой задачей по отношению к рассматриваемой постановке, так как мы не знаем наименований параметров, и нам необходимо определить, чем является вхождение переменной формулы в окружающий её сплошной текст – определением символьной переменной или обсуждением некоторых её свойств (например, типового или предельного значения).

Реализация этой возможности представляется более сложной, так как текущий механизм индексирования и поиска оперирует парами «обозначение – наименование». Нам же нужно определить, какая из окружающих символьную переменную именных групп является наименованием данного параметра.

Для решения проблем, выявленных при реализации семантического поиска математических выражений в Википедии, был разработан подход разметки для поиска по онтологии с использованием связывания формул с именными группами.

2. Разметка для поиска по онтологии

2.1. Связывание формул с именными группами. Основной задачей, решаемой модулем формульной разметки, является нахождение соответствий между формульными фрагментами документа и расположенными в их ближайшей окрестности терминологическими словосочетаниями на естественном языке. Обобщенной синтаксической моделью для терминологического словосочетания является именная группа (ИГ).

Обработка документа включает в себя следующие действия:

– выделение и анализ формульных фрагментов;

– определение связей между переменными и формульными фрагментами;

– определение связей между формульными фрагментами и именными группами;

– дополнение аннотаций Math атрибутами формульной разметки.

Модуль формульной разметки реализован на языке Java в формате плагина к текстовому процессору GATE [19], для разметки используются средства работы с аннотациями библиотеки Gate и оригинальные алгоритмы.

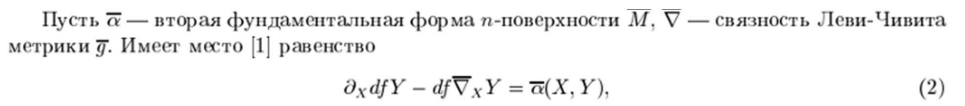
Формульная разметка применяется к XML-документу, предварительно размеченному стандартными аннотациями (Token, Sentence, Math и др.) и NLP-аннотациями (TERM, ENDS). Ключевыми аннотациями для работы алгоритма являются аннотации Math, размечающие формульные фрагменты, и аннотации TERM, соответствующие именным группам. Некоторые частные подзадачи требуют дополнительного анализа отдельных текстовых элементов, размеченных аннотациями Token, ENDS и др.

Формульным фрагментом считается любой текст, заключённый в аннотацию Math. На основе аннотаций Math строится внутренняя модель документа, содержащая набор разобранных, классифицированных, связанных между собой формульных фрагментов.

Модуль использует модифицированный алгоритм классификации математических фрагментов, применённый ранее в системе поиска формул в Википедии.

Сначала формула очищается от служебных символов языка разметки, лишних пробельных символов. Затем формульный фрагмент разбивается на элементы, разделителями считаются различные символы скобок, символы арифметических и логических операций, знаки пунктуации, пробельные символы и т. п. Полученные элементы анализируются на принадлежность к следующим специальным группам: ключевые слова (начинаются с символа “\”), нижние индексы (начинаются с символа “_”), числа и т. п. Если элемент на этом этапе не классифицирован, то с большой долей вероятности его можно считать переменной. К таким элементам дополнительно применяется проверка на соответствие правилам именования переменных – не начинается с цифры, может быть буквой греческого алфавита (например, “\alpha”), может содержать индекс. В результате все формульные фрагменты документа будут разделены на три типа: переменные, формулы и служебные (содержащие только символы разметки).

Выявленным в тексте одиночным переменным сопоставляются уникальные идентификаторы, которые будут использованы при дополнении аннотаций. Следует отметить, что алгоритм различает строчные и прописные буквы (X и x), индексированные и неиндексированные переменные (Y и Y_i), при этом варианты обозначения индекса не сказываются на идентификации переменной (Y_0 и Y_{i+1} определяются как одна и та же переменная). Для переменных, входящих в состав формульных фрагментов, строится индекс вхождения переменных в формулы. Для формул также определяются связи между фрагментами вида «содержит» и «содержится в». Служебные фрагменты не рассматриваются как не несущие смысловой нагрузки.


Пусть $\overline{\alpha}$ — вторая фундаментальная форма n-поверхности $\overline{M}$, $\overline{\nabla}$ — связность Леви-Чивита метрики $\overline{g}$. Имеет место [1] равенство

$$\partial_X df Y - df \overline{\nabla}_X Y = \overline{\alpha}(X, Y), \qquad (2)$$

Рис. 5. Фрагмент математического текста

Так, фрагмент текста на рис. 5 содержит переменные \overline{\alpha}, n, \overline{M}, \overline{\nabla}, \overline{g} и формулу, в которую входят идентифицированные переменные \overline{\alpha}, \overline{\nabla}. Переменные X, Y не присутствуют в тексте вне формул, поэтому идентификаторы им не сопоставлены.

Следующей задачей, решаемой модулем формульной разметки, является определение связей между формульными фрагментами и расположенными в их ближайшей окрестности именными группами.

Как правило, связывание имеет смысл в рамках одного предложения или некоторой его части. Для определения границ предполагаемой области связывания используются аннотации ENDS, добавленные на предыдущих этапах обработки документа. Дополнительно вводится понятие максимально допустимого расстояния (МДР) между аннотациями Math и TERM, которое определяется как наибольшее расстояние в символах между концом левой аннотации и началом правой, при котором может быть выполнено связывание. МДР является параметром, который оказывает непосредственное влияние на точность связывания и может различаться для разных коллекций документов.

В документах встречается различное взаимное расположение формул и именных групп. Формулы и именные группы могут идти последовательно друг за другом, не пересекаясь. Другой вариант – это, когда формула содержится внутри именной группы, что позволяет более чётко определить потенциальную пару для связывания. Частичное пересечение формул и именных групп не встречается, так как это противоречило бы структуре XML. Алгоритм связывания построен с учётом особенностей взаимного расположения формул и именных групп.

Для каждого формульного фрагмента с помощью средств работы с аннотациями библиотеки Gate определяется тип расположения и набор именных групп – кандидатов на связывание, из которых по заданным критериям отсеиваются неподходящие и отбираются наиболее близкие.

Если именная группа содержит формулу внутри себя, то она становится единственным кандидатом на связывание. В простом случае, когда именная группа не содержит других слов, кроме главного слова, связывание будет проведено с высокой степенью достоверности. Например, в приведённом на рис. 5 примере во втором предложении выделена ИГ «равенство $$$», где «$$$» обозначает вхождение соответствующей формулы в именную группу. В данном случае формула, обозначенная как (2), будет связана с ИГ «равенство $$$».

В более сложных случаях анализируется расстояние между формулой и главным словом именной группы (атрибуты HeadBegin, HeadEnd аннотации TERM). Если оно оказывается больше допустимого интервала между словами (3 и более символа), то считается, что формула является дополнением к основному понятию именной группы. В этом случае связывание не производится. Например, переменная \overline{g} из текста на рис. 5 не будет связана с ИГ «связность Леви-Чивита метрики $$$», так как главное слово здесь «связность Леви-Чивита». Не связываются также формулы, входящие в конструкции с дефисом, как не имеющие самостоятельного значения. Например, для фразы »вторая фундаментальная

форма n-поверхности \overline{M}» формула "n" останется несвязанной, а формула "\overline{M}" будет связана с ИГ «'вторая фундаментальная форма $$$-поверхности $$$».

Если формула не содержится внутри именной группы, задача определения набора ИГ, с которыми возможно связывание, усложняется. Сначала определяются границы области, в которой должны располагаться аннотации-кандидаты. За левую границу принимается позиция в документе, соответствующая ближайшей левой аннотации ENDS и отстоящая от начала формулы влево не более чем на МДР. Аналогично, за правую границу принимается позиция в документе, соответствующая ближайшей правой аннотации ENDS и отстоящая от конца формулы вправо не более чем на МДР. Кроме того, если формула входит в группу равенств (аннотация equationgroup), то отсчёт ведётся от начала и конца всей группы. Это необходимо для того, чтобы все уравнения группы были привязаны к одной и той же ИГ.

По границам области связывания определяются левый и правый наборы аннотаций TERM, из которых затем выбирается одна аннотация, находящаяся на минимальном расстоянии от формулы. Для повышения достоверности связывания при обнаружении распространённой в математических текстах конструкции "<формула> – <ИГ>" приоритет отдаётся правому набору (например, определение \overline{\alpha} и \overline{\nabla} на рис. 5).

Следует отметить, что алгоритм позволяет связывать формулу только с одной именной группой, но с одной и той же ИГ может быть связано более одной формулы. Это позволяет учитывать перечисления формул, относящихся по семантике к одной ИГ, но в то же время может давать некоторое количество недостоверных связываний.

На заключительном этапе построенные на внутренней модели связи переносятся в обрабатываемый документ. В аннотации Math в зависимости от типа формульного фрагмента и наличия связанных ИГ добавляются новые атрибуты. Обозначим через <variable_id> уникальный идентификатор переменной, присвоенный ей во внутренней модели документа. Отметим, что всем вхождениям какой-либо переменной в формульные фрагменты любого типа будет соответствовать один и тот же идентификатор. Переменные, встречающиеся только в составе сложных формул, идентификатора не имеют. Для формульной разметки используются следующие атрибуты:

varid=<variable_id> – идентификатор переменной, добавляется в аннотации к одиночным переменным.

vars=<variable_id1>; <variable_id2>... – список идентификаторов переменных, входящих в формулу, добавляется в аннотации к формулам.

termid=<annotation_id> – идентификатор аннотации TERM в документе, соответствующей именной группе, с которой связана формула, добавляется к переменным и формулам при наличии связи с ИГ.

В результате обработки документа, фрагмент из которого приведён на рис. 5, соответствующие фрагменту аннотации будут дополнены атрибутами формульной разметки (выделены жирным шрифтом).

Math Id=960 mode=inline, **termid=965**, tex=\overline{\alpha}, text=overline@(alpha), **varid=3,** xml:id=p10.m1

TERM Id=965 Form=вторая фундаментальная форма $$$-поверхности $$$, HeadBegin=0, HeadEnd=27

Math Id=966 mode=inline, tex=n, text=n, **varid=10**, xml:id=p10.m2

Math Id=969 mode=inline, **termid=965**, tex=\overline{M}, text=overline@(M), **varid=0**, xml:id=p10.m3

Math Id=974 mode=inline, **termid=979**, tex=\overline{\nabla}, text=overline@(nabla), **varid=8**, xml:id=p10.m4

TERM Id=979 Form=связность Леви-Чивита метрики $$$, HeadBegin=0, HeadEnd=20

Math Id=980 mode=inline, tex=\overline{g}, text=overline@(g), **varid=7**, xml:id=p10.m5

TERM Id=987 Form=равенство $$$, HeadBegin=0, HeadEnd=8

Math Id=989 mode=display, **termid=987**, tex=\partial _{{X}}dfY-df\overline{\nabla}_{{X}}Y=\overline{\alpha}(X,Y), text=(partial-differential _ X)@(d * f * Y) – d * f * (overline@(nabla)) _ X * Y = overline@(alpha) * open-interval@(X, Y), **vars=3;8**, xml:id=S0.E2.m1

2.2. Экспериментальная оценка релевантности. Выделены два возможных случая взаимного расположения формулы и именной группы: во-первых, ИГ может содержать формулу; во-вторых, элементы (формула и именная группа) могут следовать друг за другом.

В первом случае ИГ является единственным кандидатом для связывания. В простейшем случае ИГ состоит из единственного главного слова. В более сложном случае она содержит более одного слова, при этом необходимо учитывать расстояние между формулой и главным словом. Если это расстояние составляет более трёх слов, то формула считается дополнением и не связывается.

Основой анализа во втором случае является концепция максимально допустимого расстояния местонахождения в терминах позиций символов между границами математических выражений и именных групп (термов) в тексте. Для заданной пары МДР предполагается меньше длины предложения, которое содержит обе конструкции. Однако некоторые случаи анализируются специальным образом, например, такие часто встречающиеся паттерны, как «Формула – Именная группа» (со знаком тире между элементами).

Была проведена оценка релевантности и полноты связывания выявленных в тексте математических выражений и именных групп, которая основана на ручной оценке качества связывания на двух корпусах математических текстов. В качестве документов для разметки использовались статьи журнала «Известия вузов. Математика» за 1997–2009 гг.

Единственным параметром рассматриваемого метода является МДР хотя бы одного символа терма от выделенных в тексте позиций правой и левой границы математического выражения. В пределах данного расстояния производится связывание с термом, если он попал в это ограничение. На выбранном корпусе из 8 математических текстов был проведен анализ результатов связывания при изменении МДР от 15 до 40 симметрично в обе стороны. Результаты представлены в табл 1.

Для каждого заданного значения МДР на всём корпусе текстов определялись следующие параметры:

Math – количество выделенных формул;

Terms – количество выделенных ИГ;

VirOK – процент правильных связываний формул с ИГ;

NotVirOK – процент правильных несвязываний формул с ИГ (то есть констатация того факта, что математическое выражение находится в контексте, не содержащем его определения);

TotalOk – общий процент правильно обработанных формул (сумма VirOK и NotVirOK);

Табл. 1

Статистика связывания в зависимости от МДР

МДР	Math	Terms	VirOK, %	NotVirOK, %	TotalOk, %	VirBad, %	Others, %	TotalBad, %
15	1247	1357	36.33	30.47	66.80	23.90	9.30	33.20
20	1247	1357	42.34	25.50	67.84	25.66	6.50	32.16
25	1247	1357	40.98	20.69	61.67	23.02	15.32	38.33
30	1247	1357	41.38	21.49	62.87	27.83	9.30	37.13
35	1247	1357	41.86	21.01	62.87	29.03	8.10	37.13
40	1247	1357	42.02	19.65	61.67	29.67	8.66	38.33

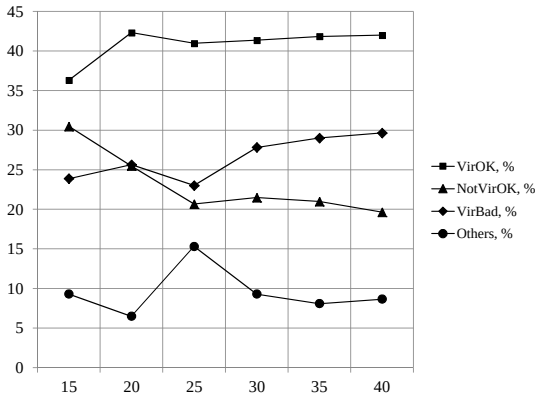


Рис. 6. Зависимость характеристик от МДР

VirBad – процент неправильных связываний формул с ИГ (из возможных кандидатов на связывание была выбрана не подходящая по семантике ИГ или произошло связывание математического выражения в контексте, не предполагающем связывания);

Others – другие ошибки связывания (отсутствие связывания там, где оно должно было быть; неправильное выделение ИГ; нераспознанные ИГ; влияющие на связывание особенности оформления текста автором);

TotalBad – общий процент неправильно обработанных формул (сумма VirBad и Others).

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

Более наглядно результаты связывания представлены на графике зависимости указанных характеристик от МДР (рис. 6), где по горизонтальной оси указано значение МДР, по вертикальной – проценты к общему количеству математических выражений в текстах.

Итоговое соотношение правильных и неправильных связываний представлено на рис. 7. Здесь максимум на графике правильных связываний совпадает с минимумом на графике неправильных связываний, это соответствует МДР=20.

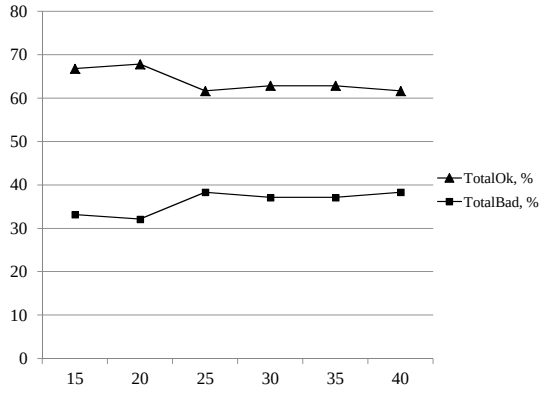


Рис. 7. Соотношение правильных и неправильных связываний

Для анализа устойчивости алгоритмов связывания был проведен расширенный анализ результатов связывания на удвоенном корпусе текстов, в том числе на нескольких документах существенно большего размера.

Результаты показали значительный разброс успешности связывания для различных документов, однако средние значения остались вполне устойчивыми. Для документов большего размера релевантность была больше средней, что является несколько неожиданным результатом. Возможно, это связано с тем, что в больших документах авторы, не связанные ограничениями по размеру, придерживались более корректного математического стиля.

В целом основным недостатком данного метода является большое количество ложных связываний выражения в контексте, не содержащем определения, что приводит к уменьшению релевантности. Для улучшения этого параметра необходимо применять дополнительные методы анализа контекста, в котором появляется математическая формула.

Заключение

В статье рассмотрены два подхода к семантическому поиску математических выражений, позволяющие выполнять запросы на поиск математических формул по текстовым наименованиям переменных, входящих в формулы.

Проведен ряд численных экспериментов на программной реализации предложенного подхода с тестовой коллекцией на основе русской Википедии. Проведена оценка релевантности и полноты связывания математических выражений и именных групп на примере двух корпусов математических текстов в зависимости от МДР. Эксперименты показали принципиальную работоспособность и неплохую устойчивость результатов предложенных подходов, а также выявили ряд проблем с релевантностью поиска.

Предложено направление дальнейшего развития предлагаемого подхода, основанное на дополнительном анализе контекста в текстах, являющихся базой поиска, а также на эвристических методах выбора именных групп для связывания и реализации механизма обучения. Выглядит также перспективным использование принципов байесовской фильтрации спама, основанного на применении наивного байесовского классификатора, для исключения незначащих слов из окружения формул [20, 21], а следовательно, для получения контекстно-зависимой вариативности МДР.

Summary

E.V. Biryaltsev, A.M. Gusenkov, O.N. Zhibrik. Some Approaches to the Markup of Scientific Documents Containing Mathematical Expressions.

This paper analyzes an approach to semantic search for mathematical expressions that would enable browsing for mathematical formulae via textual names of variables in the formulae. A method is suggested to establish relations between the textual definitions of variables, their names, and the formulae containing these variables. Two software implementations of semantic search engines are described: searching for formulae in Wikipedia articles and marking up scholarly papers for browsing via ontology. The results of experimental evaluations in the context of relevance of the search and completeness of the relations, as well as the ways of solving the main problems in the proposed approach are discussed.

Keywords: semantic search, mathematical search, formulae markup.

Литература

1. Google Scholar. – URL: http://scholar.google.com/.
2. Microsoft Academic Search. – URL: http://academic.research.microsoft.com/.
3. The Springer LaTeX Search. – URL: http://latexsearch.com/.
4. (uni)quation. – URL: http://uniquation.ru/.
5. MathWebSearch. – URL: http://search.mathweb.org/index.xhtml/.
6. *Mišutka J., Galamboš L.* Extending full text search engine for mathematical content // Sojka P. (ed.) Towards Digital Mathematics Library. Birmingham, UK, July 27th, 2008. – Brno: Masaryk University, 2008. – P. 55–67.
7. *Бирюльцев Е.В., Галимов М.Р., Жильцов Н.Г., Невзорова О.А.* Подход к семантическому поиску математических выражений в научных текстах // Материалы междунар. науч.-техн. конф. OSTIS-2012. – Минск: БГУИР, 2012. – С. 245–256.
8. *Nevzorova O., Zhiltsov N., Zaikin D., Zhibrik O., Kirillovich A., Nevzorov V., Birialtsev E.* Bringing Math to LOD: A Semantic Publishing Platform Prototype for Scientific Collections in Mathematics // The Semantic Web – ISWC'2013 (Lecture Notes in Computer Science, V. 8218). – Berlin; Heidelberg: Springer, 2013. – P. 379–394.
9. http://www.w3.org/TR/rdf-sparql-query/.
10. *Бирюльцев Е.В., Гусенков А.М.* Интеграция реляционных баз данных на основе онтологий // Учен. зап. Казан. ун-та. Сер. Физ.-матем. науки. – 2007. – Т. 149, кн. 2. – С. 13–25.
11. *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 Всерос. науч. конф. RCDL'2010. – Казань: Казан. ун-т, 2010. – С. 296–300.
12. *Невзорова О.А., Жильцов Н.Г., Бирюльцев Е.В.* Коллекции математических текстов: аннотирование и применение в поисковых задачах // Искусственный интеллект и принятие решений. – 2012. – № 3. – С. 51–62.
13. *Невзорова О.А., Жильцов Н.Г., Заикин Д.А., Жибрик О.Н., Кириллович А.В., Невзоров В.Н., Бирюльцев Е.В.* Прототип программной платформы для публикации семантических данных из математических научных коллекций в облаке LOD // Учен. зап. Казан. ун-та. Сер. Физ.-матем. науки. – 2012. – Т. 154, кн. 3. – С. 216–232.
14. *Бирюльцев Е.В., Галимов М.Р., Гусенков А.М., Жибрик О.Н.* Некоторые подходы к повышению релевантности поиска математических выражений в естественнонаучных текстах // Интеллект. Язык. Компьютер: Тр. Казан. шк. по компьютерной и когнитивной лингвистике TEL-2012. – Казань: Фэн, 2012. – Вып. 15. – С. 78–93.

15. *Бирюльцев Е.В., Гусенков А.М., Жибрик О.Н.* Поиск математических выражений в естественно-научных текстах. Экспериментальная оценка релевантности // Интеллект. Язык. Компьютер: Тр. Казан. шк. по компьютерной и когнитивной лингвистике TEL-2014. – Казань: Фэн, 2014. – Вып. 16. – С. 34–37.

16. *Бирюльцев Е.В., Гусенков А.М., Галимов М.Р.* Особенности лексико-семантической структуры наименований артефактов реляционных баз данных // Интеллект. Язык. Компьютер: Тр. Казан. шк. по компьютерной и когнитивной лингвистике TEL-2005. – Казань: Казан. гос. ун-т, 2006. – Вып. 9. – С. 4–12.

17. *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: Изд-во «Университет города Переславля», 2007. – С. 211– 216.

18. Apache Lucene. – URL: http://lucene.apache.org/.

19. GATE. – URL: http://gate.ac.uk/.

20. *Graham P.* A plan for spam. – URL: http://www.paulgraham.com/spam.html/.

21. *Guzella T.S., Caminhas W.M.* A review of machine learning approaches to spam filtering // Expert Systems with Applications. – 2009. – V. 36, No 7. – P. 10206–10222.

Поступила в редакцию
19.08.14

Биряльцев Евгений Васильевич – кандидат технических наук, заместитель директора по науке, ЗАО «Градиент», г. Казань, Россия.

E-mail: *igenbir@yandex.ru*

Гусенков Александр Михайлович – старший преподаватель кафедры технологий программирования, Казанский (Приволжский) федеральный университет, г. Казань, Россия.

E-mail: *gusenkov.a.m@gmail.com*

Жибрик Ольга Николаевна – главный программист, ООО «Градиент-технолоджи», г. Казань, Россия.

E-mail: *olgazhibrik@gmail.com*