УДК 004.91

СЕМАНТИЧЕСКИЙ АНАЛИЗ ДОКУМЕНТОВ В СИСТЕМЕ УПРАВЛЕНИЯ ЦИФРОВЫМИ НАУЧНЫМИ КОЛЛЕКЦИЯМИ

Ш.М. Хайдаров

Институт математики и механики им. Н.И. Лобачевского Казанского (Приволжского) федерального университета
15jkeee @gmail.com

Аннотация

Предложены методы семантического анализа документов в системе управления цифровыми научными коллекциями, в том числе электронными научными журналами. Рассмотрены методы обработки документов, содержащих математические формулы, а также способы конвертации этих документов из формата OpenXML в формат TeX. Разработан алгоритм поиска по формулам в коллекциях математических документов, хранящихся в формате OpenXML. Алгоритм реализован в виде онлайн-сервиса на платформе science.tatarstan.

***Ключевые слова*:** семантический анализ, издательские системы.

ВВЕДЕНИЕ

В настоящее время объемы электронных данных возрастают с колоссальной скоростью, а научная деятельность неразрывно связана с использованием информационно-коммуникационных технологий. Электронное представление научных документов вошло в каноны научной деятельности; уже многие журналы перешли на электронный способ приема материалов, практически все журналы выставляют опубликованные работы в открытый доступ в интернет (с временным эмбарго либо непосредственно после опубликования соответствующего номера или выпуска) (см., например, [1]).

В «Издательской Вселенной» – множество «жителей»: авторы, рецензенты, редакторы, издатели, библиотеки, читатели, университеты и научное сообщество. Для упорядочения жизненного цикла создания публикаций, а также последующего их распространения и хранения уже существуют различные систе-

мы, распределяющие как роли взаимодействующих пользователей, так и позволяющие группировать различные издания и хранить публикации по определенным нормам и правилам (см., например, [2, 3]). Подобные системы значительно упрощают публикационный процесс, перестраивают классические этапы получения и передачи статей рецензентам и дальнейшей их обработки. В результате появляется возможность автоматизировать многие из этих процессов. При этом сами такие системы нуждаются в дополнительных сервисах, позволяющих оптимизировать процессы, поддерживаемые этими. Одна из имеющихся проблем заключается в различиях форматов представления (оформления) публикаций, используемых в тех или иных журналах. Отличаются также методы набора текстов и форматы их последующего хранения, не говоря уже о способе воспроизведения математических формул. Мало того, что формулы можно воспроизвести различными средствами, с помощью этих средств одни и те же формулы могут быть представлены разнообразными способами. Таким образом, в большинстве случаев ошибки при наборе формул в материалах, представляемых к публикации, однотипны и связаны, прежде всего, с оформлением математических выражений по шаблону, применяемому в конкретных журнале или издании.

Настоящая статья посвящена решению одной из таких проблем, когда различия в форматах представления и хранения документов создают значительные неудобства как авторам научных статей, так и редколлегиям журналов, принимающих эти материалы к публикации и использующих определенные форматы их представления (см., например, [4]). Таким образом, речь идет о системах, позволяющих извлекать данные из одних форматов и преобразовывать их в другой. В частности, это касается документов, созданных с использованием офисных пакетов.

ОБЩЕПРИНЯТЫЕ ФОРМАТЫ ХРАНЕНИЯ ДАННЫХ

Нотация TeX. На момент создания в 1978 году Дональдом Кнутом цифровой системы TeX для представления математических документов она не была ни первой, ни единственной подобной системой. Быстро завоевав признание, сегодня она стала основным инструментом набора математических текстов и важнейшим средством коммуникации научного сообщества. В системе TeX пользо-

ватель задает текст и его структуру, а система на основе выбранного пользователем шаблона самостоятельно форматирует документ, заменяя при этом дизайнера и верстальщика. На данный момент TeX является основным инструментом представления математических формул. Однако, в виду сложности представления документов, созданных в TeX, был разработан другой метод представления математической информации в Вебе. Этим форматом стал ***язык разметки MathML,*** созданный в 1999 году консорциумом W3C как способ представления математических формул в Вебе. В результате MathML значительно поменял принципы организации и управления электронными публикациями по математике.

Офисные пакеты. При подготовке научных документов все чаще используют офисные пакеты. Их широкое распространение обусловлено простотой ввода и наглядностью конечного (печатного) варианта документов (статей, докладов, книг и пр.). Наиболее распространённый формат хранения офисных документов **.doc* устарел, и на его замену пришел формат OpenXML, основанный на открытых семантических технологиях и имеющий расширение **.docx*. Ниже описаны методы машинного извлечения данных из документов в формате OpenXML.

ИЗВЛЕЧЕНИЕ ДАННЫХ ИЗ ДОКУМЕНТОВ В ФОРМАТЕ OpenXML

Формат OpenXML основан на открытых технологиях семантического Веба и представляет собой сжатый ZIP-контейнер, содержащий как разметку документа, так и вспомогательные его части, такие, как шрифты и стили (см., например, [5, 6]). Основной скелет документа в структурированном виде хранится в файле *«word/document.xml»*. Здесь содержится текст документа с его форматированием (см. фрагмент кода на рис. 1, 2) и всеми его объектами или ссылками на них, например, изображения хранятся в отдельном каталоге, однако их позиция и свойства описаны в данном файле.

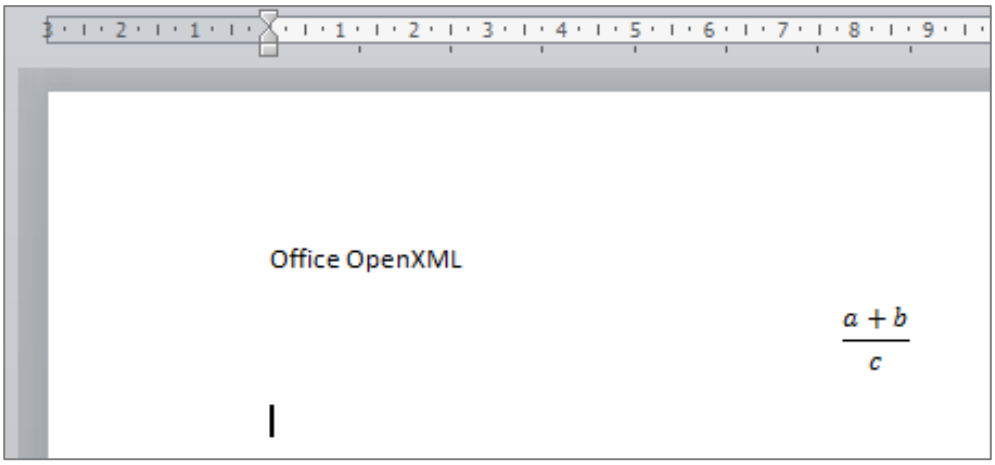

Рис. 1. Представление кода

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<w:document     xmlns:wpc="
http://schemas.microsoft.com/office/word/2010/wordprocessingCanvas"
    <w:body> <!--начало документа-->
        <w:p w:rsidR="00AF6381" w:rsidRPr="00DA088A" w:rsidRDefault=
        "00B300E9"> <!--Начало абзаца-->
        <w:pPr> <!--Свойства абзаца-->
            <w:rPr>
                <w:rFonts w:ascii="Calibri" w:eastAsia="Calibri"/>
            </w:rPr>
        </w:pPr>
        <w:r w:rsidRPr="00DA088A">
            <w:t>Office OpenXML</w:t> <!--Текст.-->
        </w:r>
    </w:p>
    <m:oMath><!--Тег начала формул, пространство имён OpenMathML-->
        <m:f> <!--Тег дроби-->
            <m:fPr> <!--Свойства дроби и стили текста-->
            </m:fPr>
            <m:num> <!--Тег числителя-->
                <m:t>a+b</m:t>
            </m:num>
            <m:den> <!--Тег знаменателя-->
                <m:t>c</m:t>
            </m:den>
        </m:f>
    </m:oMath>
    </w:body>
</w:document>
```

Рис. 2. Фрагмент кода в OpenXML

Извлечение элементов документа. При обработке документа сначала необходимо извлечь из архива XML-файлы с метаданными. В случае с текстом достаточно из архива получить файл *«word/document.xml».* Если стоит задача лишь выделения обычного текста без всякого форматирования, то достаточно вернуть значения тегов *«w:t»*. Однако, чтобы применить выбранные стили, необходимо обработать атрибуты тегов [7, 8]. На рис. 3 показан фрагмент шаблона XSLT, обрабатывающего текст с основными стилями, на примере перевода

его в TeX-нотацию.

```
22  <xsl:template match="w:p">
23      <xsl:apply-templates/>
24      <xsl:if test="position()!=last()"><xsl:text>
25
26  </xsl:text>
27      </xsl:if>
28  </xsl:template>
29
30  <xsl:template match="w:r">
31      <xsl:if test="w:footnoteReference">
32          <xsl:text>\footnote{</xsl:text>
33          <xsl:call-template name="footnote">
34              <xsl:with-param name="fid">
35                  <xsl:value-of select="//@w:id"/>
36              </xsl:with-param>
37          </xsl:call-template>
38          <xsl:text>}</xsl:text>
39      </xsl:if>
40      <xsl:if test="w:rPr/w:b">
41          <xsl:text>\textbf{</xsl:text>
42      </xsl:if>
43          <xsl:call-template name="pastb"/>
44      <xsl:if test="w:rPr/w:b">
45          <xsl:text>}</xsl:text>
46      </xsl:if>
47  </xsl:template>
48
49  <xsl:template name="pastb">
50      <xsl:if test="w:rPr/w:i">
51          <xsl:text>\textit{</xsl:text>
52      </xsl:if>
53      <xsl:call-template name="pasti"/>
54      <xsl:if test="w:rPr/w:i">
55          <xsl:text>}</xsl:text>
56      </xsl:if>
57  </xsl:template>
```

Рис. 3. Фрагмент шаблона, обрабатывающего текст в OpenXML

Что касается изображений, графиков и элементов из других приложений, то они хранятся в отдельной папке в бинарном виде, а в приведенном xml-файле содержится лишь ссылка, характеризующая положение этой папки, применен-ные фильтры и др. (см. рис. 4).

```
<a:graphic xmlns:a="http://schemas.openxmlformats.org/drawingml/2006/main">
    <a:graphicData uri="http://schemas.openxmlformats.org/drawingml/2006/picture">
    <pic:pic xmlns:pic="http://schemas.openxmlformats.org/drawingml/2006/picture">
        <pic:nvPicPr>
            <pic:cNvPr id="0" name="P1030639.JPG"/>
            <pic:cNvPicPr/>
        </pic:nvPicPr>
        <pic:blipFill>
            <a:blip r:embed="rId5" cstate="print">
                <a:extLst>
                    <a:ext uri="{28A0092B-C50C-407E-A947-70E740481C1C}">
                        <a14:useLocalDpi xmlns:a14=
        "http://schemas.microsoft.com/office/drawing/2010/main" val="0"/>
                    </a:ext>
                </a:extLst>
            </a:blip>
            <a:stretch>
                <a:fillRect/>
            </a:stretch>
        </pic:blipFill>
        <pic:spPr>
            <a:xfrm>
                <a:off x="0" y="0"/>
                <a:ext cx="5940425" cy="4455160"/>
            </a:xfrm>
            <a:prstGeom prst="rect">
                <a:avLst/>
            </a:prstGeom>
        </pic:spPr>
    </pic:pic>
    </a:graphicData>
</a:graphic>
```

Рис. 4. Фрагмент OpenXML со ссылкой на изображение

Извлечение математических формул. Одним из подмножеств формата OpenXML является Office Math Markup Language (OMML) – язык математической разметки, который встроен в WordprocessingML (кроме того, он может быть использован в SpreadsheetML и PresentationML). Во фрагменте, приведенном на рис. 2, показано, как выглядит простая дробь в формате OMML, это представление хранится в окружении тега «*m:oMath*».

Важно отметить следующее обстоятельство: вследствие того, что OMML является подмножеством основного языка разметки WordprocessingML, как у любого элемента во всем документе, у каждого элемента имеется свое форматирование, что позволяет применить конкретные стили для любой части математического текста.

Хотя языки разметки MathML и OMML весьма схожи между собой, между ними имеется существенное различие. Организация математических элементов в MathML происходит по положению, а в OMML – по точному названию (см. рис.

5 для сравнения, как выглядит уже знакомая формула).

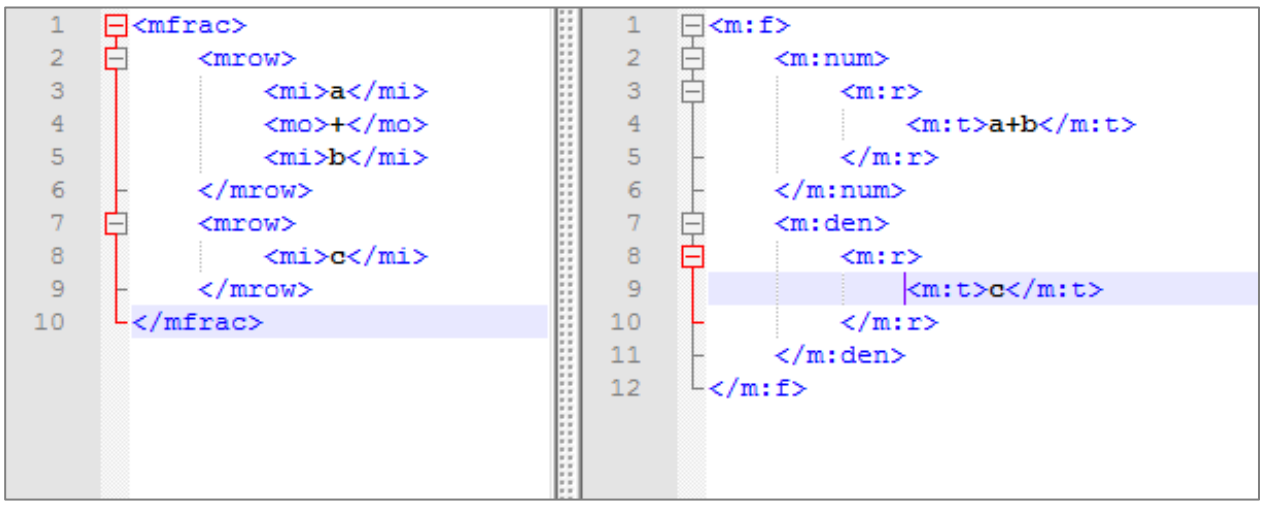

Рис. 5. Различие представления в MathML и OMML

Таким образом, в формате MathML каждый элемент (число, переменная или функция) является самостоятельным, а в OMML все они группируются по схожим характеристикам (например, все функции записываются в виде текста). По этой причине возникают неточности в переводе при использовании данных шаблонов. Например, при переводе OMML в MathML все элементарные функции, которые в MathML имеют свои теги, записываются в виде математического текста (сравните: <cos/> в «чистом» MathML и <mml:mi mathvariant="normal">cos</mml:mi> — в виде, переведенном из OMML). В MathML отсутствует поэлементное форматирование математического текста, что необходимо учитывать при реализации поиска по формульным фрагментам.

В пакете Microsoft Office содержатся средства для работы с OMML, реализованные в виде XSLT-шаблонов. Эти шаблоны позволяют получать из документа формата OMML документ формата MathML и обратно. В Microsoft Office версии 2010 эти функции выполняют шаблоны «MML2OMML.xsl» и «OMML2MML.xsl», расположенные в директории «%ProgramFiles%\Microsoft Office\Office14». Данные шаблоны предполагают, что используются пространства имен

«xmlns:mml="http://www.w3.org/1998/Math/MathML"»

и

«xmlns:m="http://schemas.openxmlformats.org/officeDocument/2006/math"».

Таким образом, теги из других пространств имен будут просто пропущены.

Чтобы их не пропустить, достаточно в таблицы стилей добавить ссылки на соответствующие пространства. Например, для текстовой разметки это *«xmlns:w="http://schemas.openxmlformats.org/wordprocessingml/2006/main"»* и соответственно шаблон, обрабатывающий эти теги. Например, в случае перевода в TeX-нотацию этот шаблон будет следующим (см. рис. 6):

```
23  <xsl:template match="w:body">
24  <xsl:text>\begin{document}
25  </xsl:text>
26   <xsl:apply-templates/>
27  <xsl:text>
28  \end{document}</xsl:text>
29  </xsl:template>
30
31  <xsl:template match="w:p">
32      <xsl:apply-templates/>
33      <xsl:if test="position()!=last()">
34          <xsl:text>
35
36  </xsl:text></xsl:if>
37  </xsl:template>
```

Рис. 6. Фрагмент шаблона, обрабатывающего текст

Для конвертации в TeX-формат разработана таблица стилей «d2t.xsl», работа которой сводится к извлечению файла с разметкой основного каркаса docx-документа с последующей обработкой этого файла XSLT-процессором.

На рис. 7 приведен фрагмент таблицы стилей «d2t.xsl», которая преобразует дроби. Код записи дроби в формате OpenXML приведен на рис. 2. Таким образом, в исходном файле обрабатываются только те теги, которые описаны в таблицах стилей.

```
<xsl:template match="m:f"> <!--если исходный файл содержит данный тег,
то будет выполнен этот шаблон-->
    <xsl:variable name="sLowerCaseType" select=
    "translate(m:fPr[last()]/m:type/@m:val,
    'ABCDEFGHIJKLMNOPQRSTUVWXYZ', 'abcdefghijklmnopqrstuvwxyz')" />
    <!--Создает переменную в которой написано свойство дроби, и
    переводит в нижний регистр-->
    <xsl:choose> <!---->
        <xsl:when test="$sLowerCaseType=''"> <!--если свойство пу-стое,
        то по-умолчанию выбирается обычная дробь-->
            <xsl:text>\frac{</xsl:text> <!--Добавляется текст-->
            <xsl:apply-templates select="m:num[1]" /> <!--числитель-->
            <xsl:text>}{</xsl:text>
            <xsl:apply-templates select="m:den[1]" /> <!--знаменатель-->
            <xsl:text>}</xsl:text>
        </xsl:when>
        <xsl:when test="$sLowerCaseType='lin'">
            <xsl:apply-templates select="m:num[1]" />
            <xsl:text>\!/</xsl:text>
            <xsl:apply-templates select="m:den[1]" />
        </xsl:when>
    </xsl:choose>
</xsl:template>
```

Рис. 7. Шаблон, преобразующий дроби из формата OMML в формат TeX

Конвертер поддерживает основное форматирование текста, формулы: все математические функции, поддерживаемые Word (дроби, радикалы, тригонометрические функции, пределы и пр.), операторы (интегралы, ряды и пр.), матрицы и системы уравнений, диакритические знаки, греческий алфавит и др.

Отметим, что нотация TeX не содержит тегов для всех символов кодировки Unicode. Так как OpenXML для записи символов использует Unicode, ряд символов в TeX отсутствует. Поэтому при конвертации необходимо использовать дополнительные программные средства.

На следующих рисунках приведен пример конвертации из OpenXML в TeX.

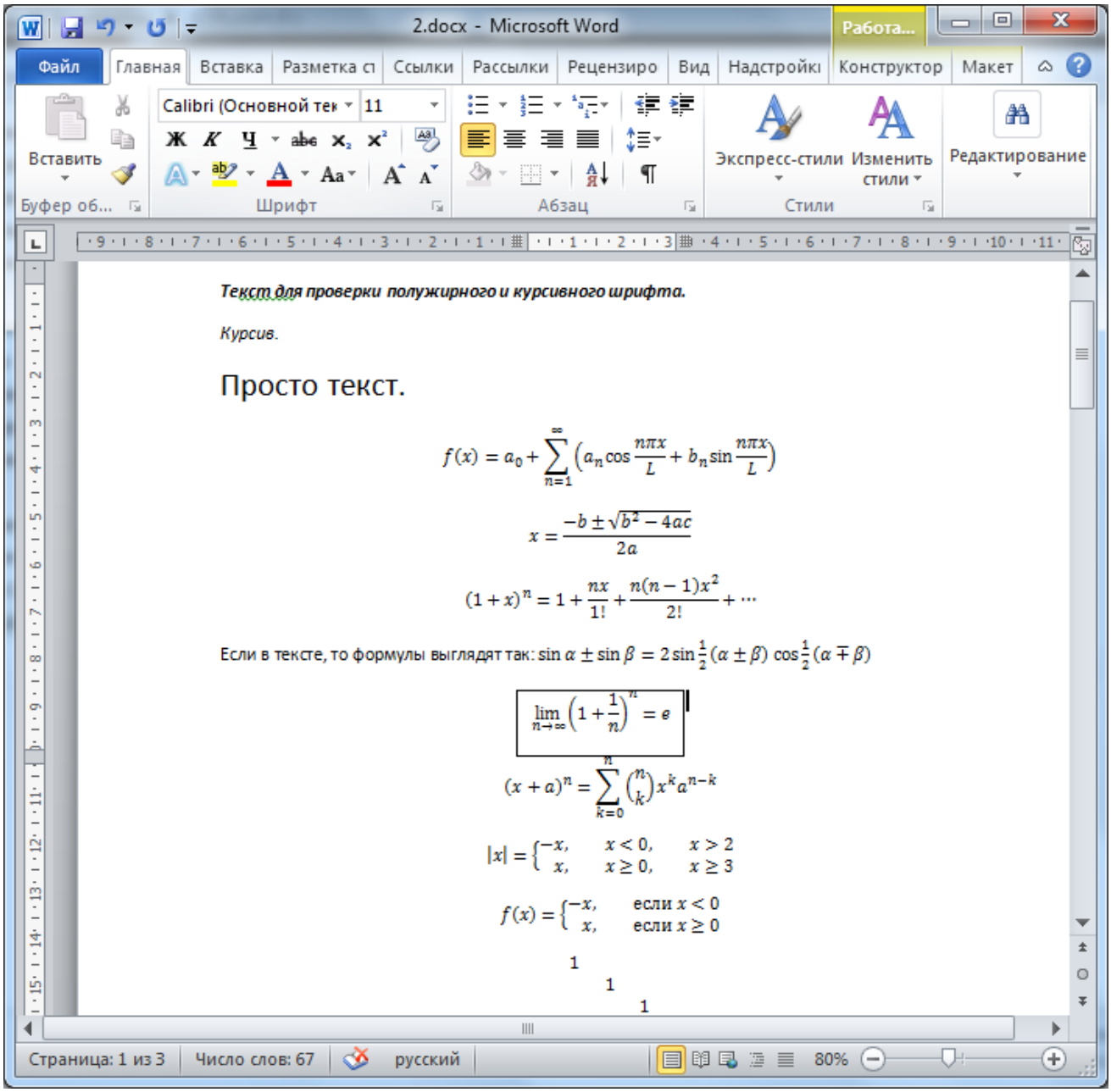

Рис. 8. Исходный документ в формате OpenXML

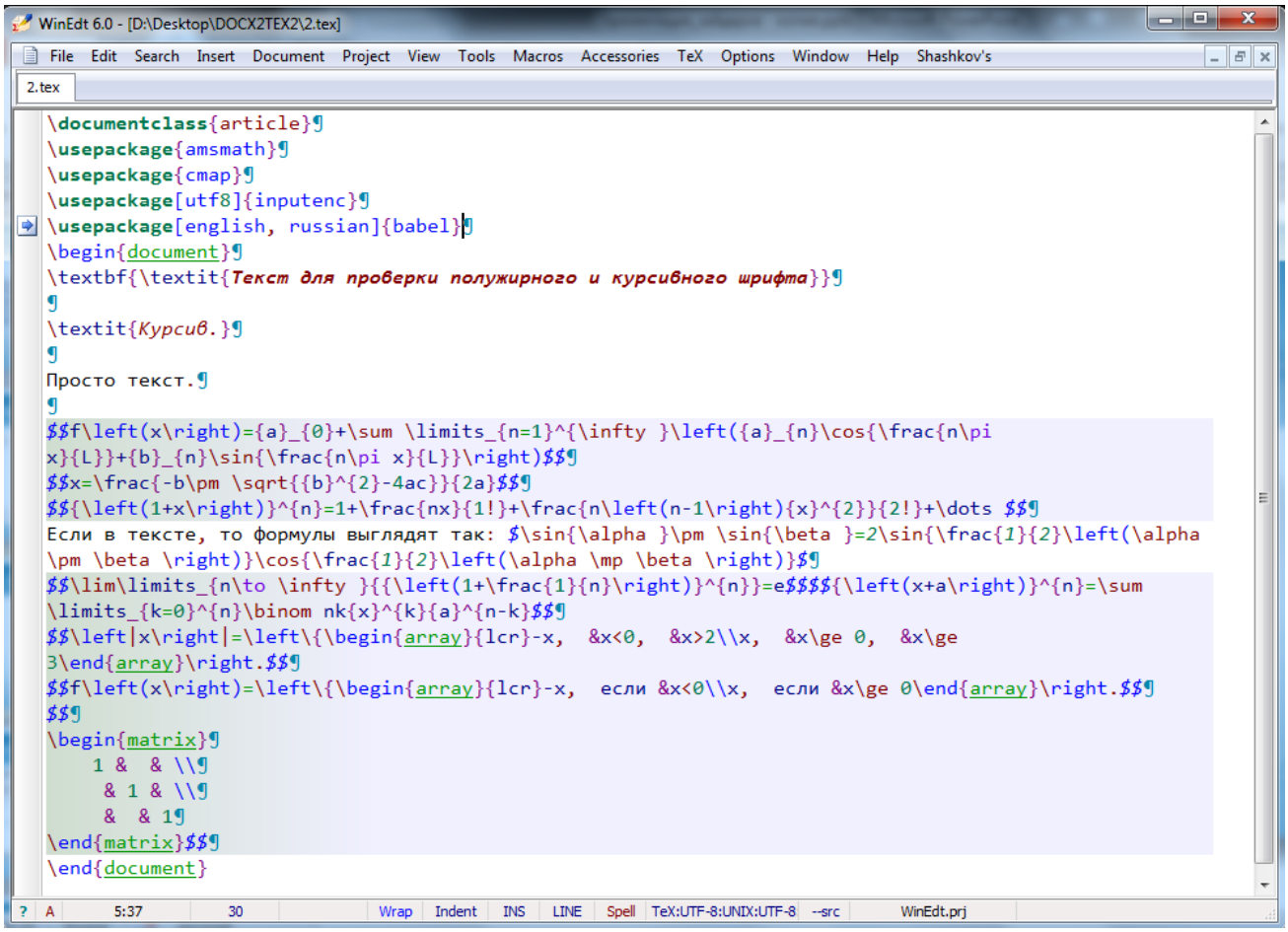

Рис. 9. TeX-код, полученный в результате конвертации

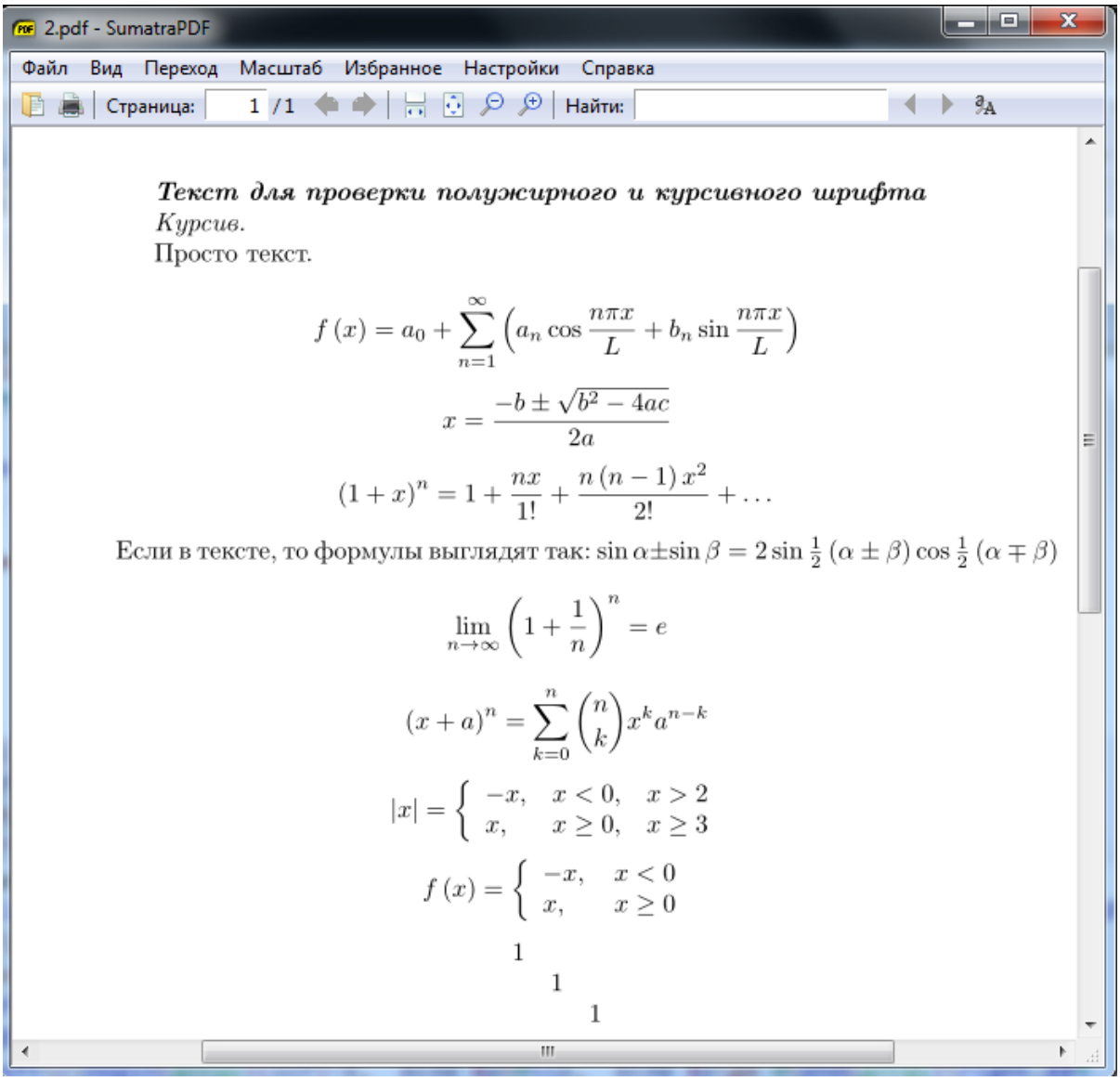

Рис. 10. Документ, полученный в результате работы конвертера

Алгоритм реализован в виде онлайн-сервиса на платформе science.tatarstan [9].

ПОИСК ПО МАТЕМАТИЧЕСКИМ ФОРМУЛАМ

С развитием веб-технологий в интернете размещается все больше результатов научных работ (это относится и к математическим исследованиям), одновременно все более важным становится развитие поисковых инструментов для анализа научного контента. Традиционный поиск, организованный по ключевым словам, как правило, не способен находить математические формулы, интере-

сующие пользователя. Правда, иногда поисковики выдают требуемый результат, но чаще всего это происходит из-за удачно подобранных ключевых слов и словосочетаний. Тем не менее, основная информация в математических текстах содержится именно в формулах. Поэтому актуальным является разработка методов, способных анализировать математический контент и осуществлять поиск по формулам (см., например, [8]).

Среди подходов к решению задачи поиска по математическим выражениям выделяют синтактический поиск [2] и семантический поиск, основанный на онтологиях предметных областей [10–12]. Последний, в отличие от синтаксического поиска, разбивает документ на семантические области, выполняет семантическую разметку документа и создает онтологические связи [13]. Это позволяет вести поиск в различных фрагментах текста (определениях, теоремах, доказательствах и т. п.). После индексирования коллекции математических документов для каждого документа формируется семантическое описание (например, в форматах STeX и RDF (см. [12, 13]).

Канонизация представления. Важный этап обработки математических текстов – представление поисковых запросов в едином формате. Для этого необходима конвертация в этот формат документов, представленных в других форматах. При реализации названного этапа преимуществами обладают языки семантической разметки, такие, как MathML (например, [14, 15]). Эффективность использования MathML в качестве формата хранения документов подтверждена практикой его использования в журнале Lobachevskii Journal of Mathematics [2], [16]. Отметим, что MathML реализован в двух вариантах: Presentation MathML и Content MathML [17]. Использование Presentation MathML при обработке математического контента не эффективно в силу следующих причин [12, 18, 19]:

1. *Многообразие математических терминов и обозначений*: в разных разделах математики термины могут иметь различный смысл, то же касается и обозначений, например, формула $\frac{n!}{k!(n-k)!}$ может обозначаться как $\binom{n}{k}$, ${}_k\mathrm{C}^n$, C_n^k, C_k^n.

2. *Выбор обозначений:* при переобозначении переменных смысл формулы не изменяется (например, результаты поиска формул $\int f(x)dx$ и $\int g(y)dy$ должны совпадать).

3. *Различные способы записи математических выражений*: например, степени можно записать словами, надстрочным знаком и т. п.

Часть указанных проблем решают языки разметки содержательного уровня, такие, как Content MathML и OpenMath.

Для конвертации из Content MathML в Presentation MathML можно использовать XSLT-преобразование (см., например, [20]). Обратная же задача является одной из проблем искусственного интеллекта (см. [21]). Отметим в связи с этим Java-библиотеку SnuggleTeX [22], предназначенную для конвертации фрагментов математических формул в различных формах представления – из TeX в Presentation MathML, из Presentation MathML – в Content MathML, а также в формат Maxima. Наибольший интерес представляет преобразование в Content MathML: для этого используются как средства XSLT, так и средства Java, однако поддержка ограничивается только арифметическими операциями и функциями. Кроме того, Content MathML используется как формат хранения документов в пакете Wolfram Mathematica [23]. Обусловлено это тем, что этот пакет сохраняет структуру введенных данных, причем при вводе данные проходят валидацию.

***Подход к решению проблемы поиска по формулам*.** В настоящей работе предложен метод структурного поиска по математическим выражениям, хранящимся в документах в форматах OpenXML и TeX. Использован язык программирования PHP для возможности интеграции с издательской платформой OJS . Алгоритм поиска по математическим выражениям аналогичен подходу, реализованному в переносимой коллекции электронных математических документов журнала Lobachevskii Journal of Mathematics (см. [2], [15]). Программа поиска включает два блока: первый индексирует математические выражения в коллекции, второй представляет пользовательский интерфейс поиска, обрабатывающий запрос и выдающий найденные результаты. Создание индексируемого

файла позволяет увеличить скорость поиска за счет ограничения времени обработки документов.

Индексирование документов в формате OpenXML проводится по следующей схеме:

- в выбранной директории создается список документов с расширением *.docx;*
- поскольку каждый файл *.docx* является архивом [5], [6], производится его распаковка и извлекается файл *«word/document.xml»;* на этом шаге производится анализ этого файла и исключаются документы без математических формул (по наличию тега «o:Math»);
- с использованием таблицы стилей *«OMML2MML.XSL»,* входящей в состав Microsoft Office, документ конвертируется в формат MathML; дополнительно в таблицу стилей вносятся изменения для унификации записи (например, пределов интегрирования или суммирования, см. рис. 11);
- в полученном документе удаляются упоминания пространств имен, атрибуты и параметры форматирования;
- создается xml-файл (индекс-файл), в котором сгруппированы все формулы документов директории (рис. 12).

```
<xsl:template match="m:nary">
   <mml:munderover>
     <xsl:call-template name="OutputNaryMo">
       <xsl:with-param name="ndCur" select="." />
       <xsl:with-param name="fGrow" select="$fGrow" />
     </xsl:call-template>
     <mml:mrow>
       <xsl:apply-templates select="m:sub[1]" />
     </mml:mrow>
     <mml:mrow>
       <xsl:apply-templates select="m:sup[1]" />
     </mml:mrow>
   </mml:munderover>
   <mml:mrow>
     <xsl:apply-templates select="m:e[1]" />
   </mml:mrow>
 </mml:mrow>
</xsl:template>
```

Рис. 11. Изменения, вносимые в файл OMML2MML.xsl, для унификации записи операторов

```
<?xml version="1.0"?>
<files>
    <file name="files//2.docx">
        <math>Формула 1..</math>
        <math>Формула 2..</math>
        <math>Формула 3..</math>
    </file>
    <file name="files//3.docx">
        <math>Формула 1..</math>
        <math>Формула 2..</math>
    </file>
</files>
```

Рис. 12. Фрагмент xml-файла, содержащего сгруппированные формулы

Основные шаги алгоритма таковы (см. рис. 13):

- вводится поисковый запрос в TeX-нотации, для его отображения в браузере и дальнейшей обработки используются библиотеки MathJax [24, 25] и JQuery [26], средствами которых генерируется код запроса на языке MathML;
- в этом коде удаляются теги пространств имен, атрибутов и параметров форматирования;
- записывается регулярное выражение, позволяющее учесть различия в записях переменных в одной и той же формуле в различных документах коллекции, где осуществляется поиск (см. рис. 14, 15);
- выполняется поиск совпадений в индекс-файле, формируется набор гиперссылок на документы коллекции, содержащие искомую формулу (пример см. на рис. 16).

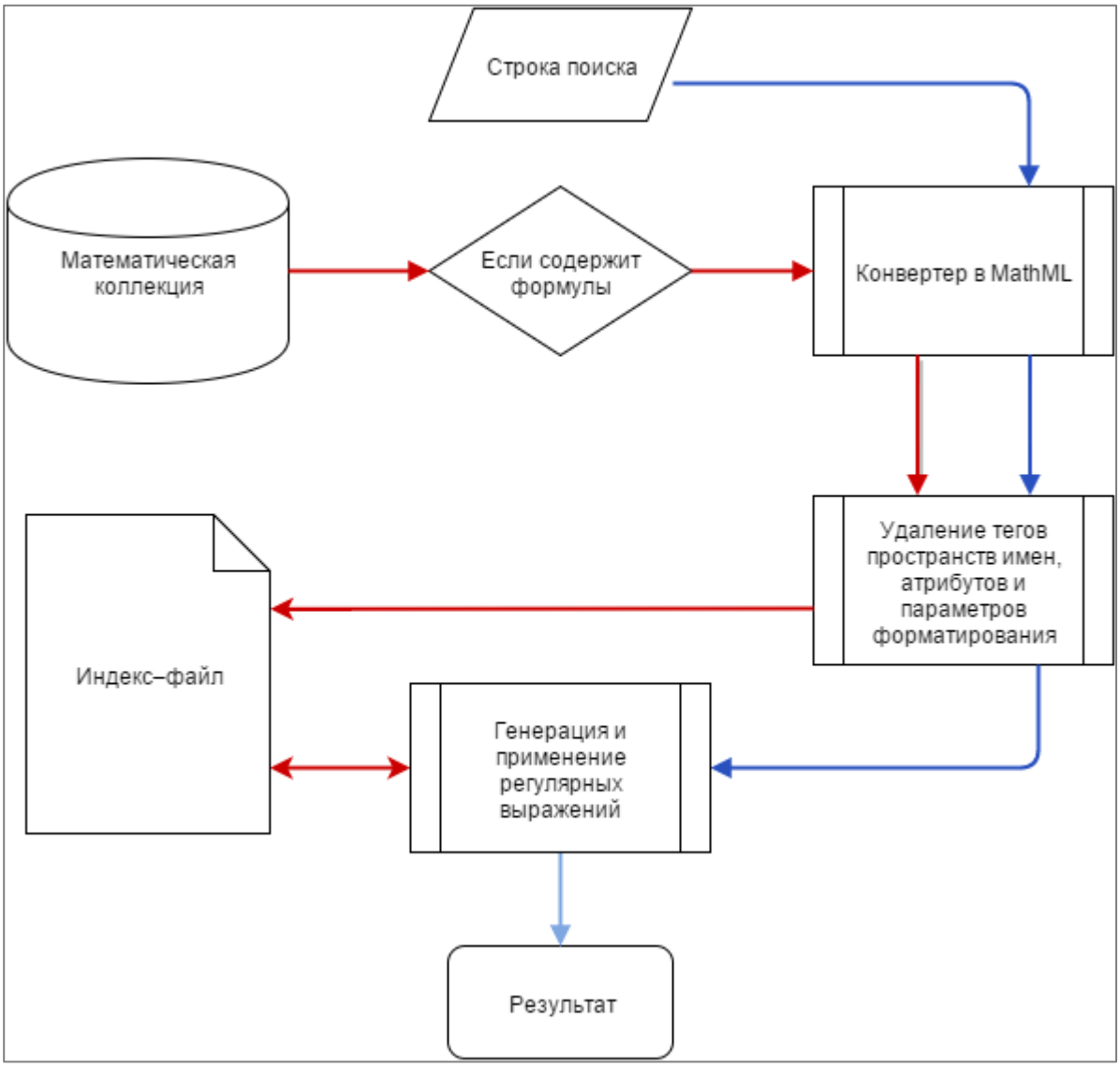

Рис. 13. Алгоритм работы программы поиска по формулам

```
21  //Эквивалентность.
22  preg_match_all("/(?<=(<mi>))$excludedvars(?=(<\/mi>))/",$query,$vars);
23  $vars=array_count_values($vars[0]);
24  foreach($vars as $key=>$var){
25      if($var>1){
26          $query=preg_replace("/(?<=(<mi>))$key(?=(<\/mi>))/","(.)",$query,1);
27          $query=str_replace("<mi>$key</mi>","<mi>\\1</mi>",$query);
28      }else $query=preg_replace("/(?<=(<mi>))$key(?=(<\/mi>))/","$excludedvars",$query,1);
29  }
30
```

Рис. 14. Фрагмент кода для учета различий в записях переменных в одной и той же формуле

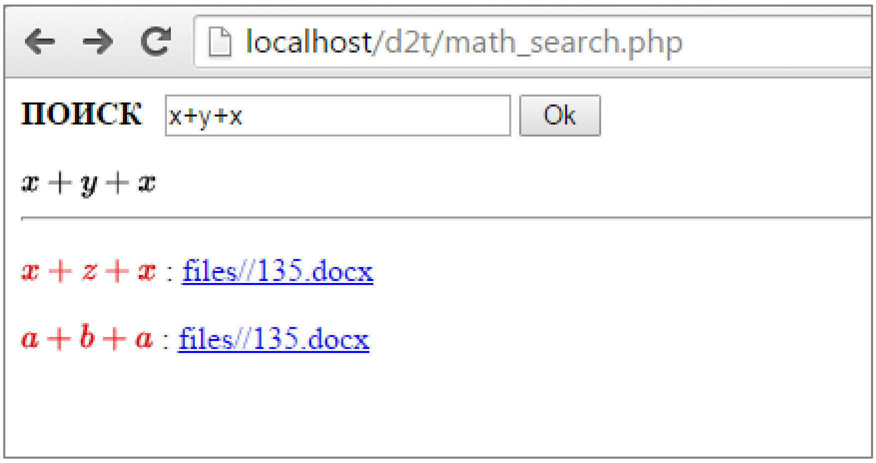

Рис. 15. Иллюстрация обработки свободных переменных

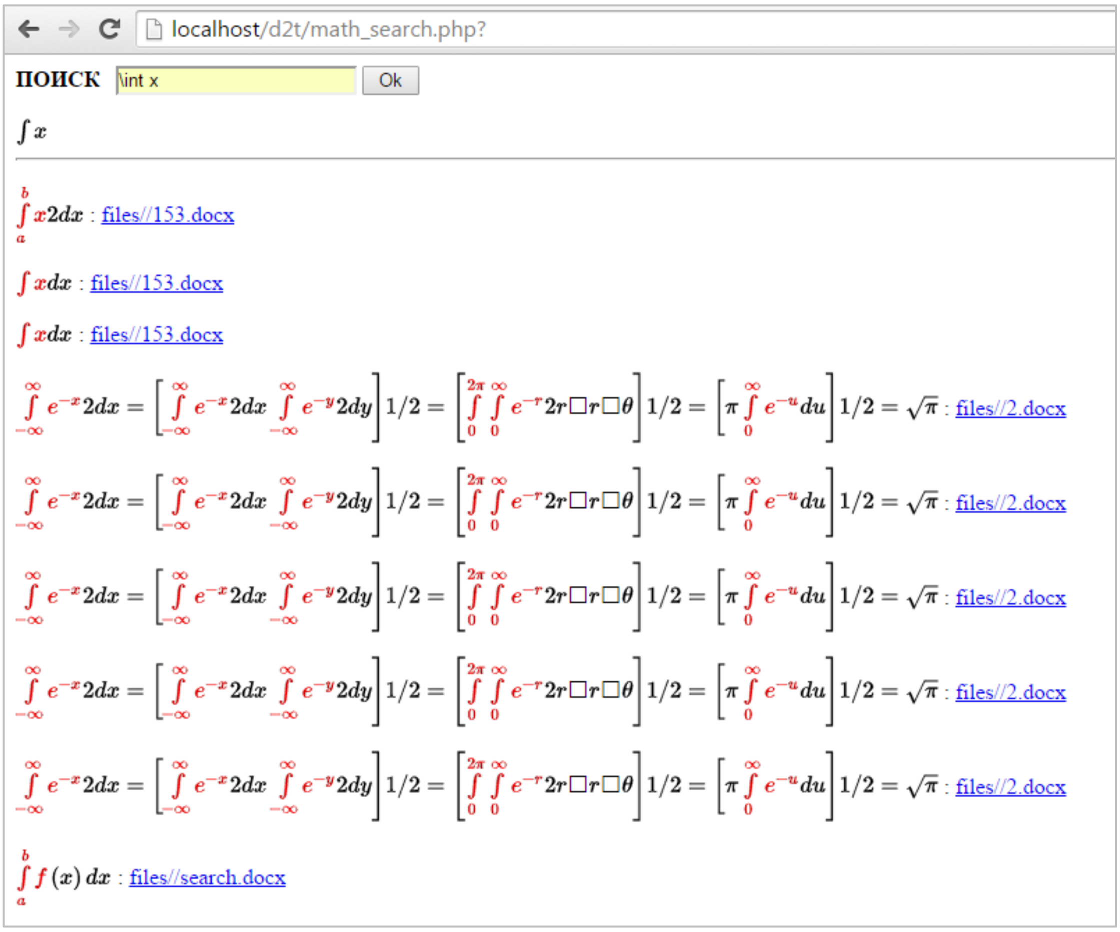

Рис. 16. Отображение результатов поиска по запросу \int x

ЗАКЛЮЧЕНИЕ

Наиболее эффективными в организации поиска представляются подходы, основанные на семантических языках содержательного уровня, таких, как Content MathML. Однако ввиду отсутствия программных средств, позволяющих качественно и полноценно извлекать код из форматов хранения презентационного уровня, возникают трудности в обработке таких выражений. Тем не менее, все форматы, описанные выше, можно привести к общему виду, используя такие технологии, как XSLT и JavaScript. Изложенный подход позволил реализовать поиск по документам в формате OpenXML. Кроме того, частично решена проблема поиска по математическим формулам.

БЛАГОДАРНОСТИ

Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда, проект 14-03-12004, а также Российского фонда фундаментальных исследований и Правительства Республики Татарстан в рамках научного проекта № 15-47-02472.

СПИСОК ЛИТЕРАТУРЫ

1. *Елизаров А.М., Липачев Е.К., Хохлов Ю.Е.* Семантические методы структурирования математического контента, обеспечивающие расширенную поисковую функциональность // Информационное общество. 2013. № 1–2. С. 83-92.

2. *Елизаров А.М., Липачев Е.К., Малахальцев М.А.* Веб-технологии в работе электронного математического журнала Lobachevskii Journal of Mathematics // В сборнике: Научный сервис в сети Интернет: многоядерный компьютерный мир. 15 лет РФФИ. Труды Всероссийской научной конференции. Московский государственный университет им. М.В. Ломоносова; Южный федеральный университет; Институт вычислительной математики РАН, г. Москва. 2007. С. 355-356.

3. *Елизаров А.М., Зуев Д.С., Липачёв Е.К.* Информационные системы управления электронными научными журналами // Научно-техническая информация. Серия 1. Организация и методика информационной работы. 2014. № 3. С. 31-38.

4. *Хайдаров Ш.М.* Методы управления математическим контентом в информационных издательских системах // Тр. Математического центра им. Н.И.

Лобачев-ского. Материалы 14-й Всерос. Молодежной школы-конференции «Лобачевские чтения–2015» (Казань, 22–27 октября 2015 года). Казань. 2015. С. 162-165.

5. *Воутер В.В.* Open XML – Кратко и доступно. Open XML Technical Evangelist, Microsoft, 2007. 101 с.

6. *Standard ECMA-376: Office Open XML File Formats* [Электронный ресурс] URL: http://www.ecmainternational.org/publications/standards/Ecma-376.htm.

7. *Липачёв Е.К., Хайдаров Ш.М.* Система сервисов преобразования электронных математических документов на основе облачных технологий // Труды Математического центра им. Н.И. Лобачевского. Казань. 2013. Т. 47. С. 109-110.

8. *Елизаров А.М., Липачёв Е.К., Невзорова О.А., Соловьев В.Д.* Методы и средства семантического структурирования электронных математических документов // Доклады Академии наук. 2014. Т. 457. № 6. С. 642-645.

9. *Ахметов Д.Ю., Герасимов А.Н., Грачев А.О., Елизаров А.М., Липачёв Е.К.* Облачная платформа поддержки электронных научных изданий // Учёные записки Института социальных и гуманитарных знаний. 2014. № 1 (12), ч.1. С. 13-19.

10. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A., Solovyev V.D., Zhiltsov N.G.* Mathematical knowledge representation: semantic models and formalisms // Lobachevskii Journal of Mathematics. 2014. V. 35. No 4. P. 348-354.

11. *Биряльцев Е.В., Елизаров А.М., Жильцов Н.Г., Липачёв Е.К., Невзорова О.А., Соловьев В.Д.* Методы анализа семантических данных математических электронных коллекций // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2014. № 4. С. 12-17.

12. *Биряльцев Е.В., Галимов М.Р., Жильцов Н.Г., Невзорова О.А.* Подход к семантическому поиску математических выражений // OSTIS-2012. 2012. С. 245-256.

13. *Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E.* OntoMathPRO Ontology: A Linked Data Hub for Mathematics // Knowledge Engineering and the Se-mantic Web Communications in Computer and Information Science. 2014. V. 468. P. 105-119.

14. *Веселаго В.Г., Елизаров А.М., Липачев Е.К., Малахальцев М.А.* Формирование и поддержка физико-математических электронных научных изданий: переход на технологии Семантического Веба // Научно-исследовательский институт математики и механики им. Н.Г. Чеботарева Казанского государственного университета. 2003–2007 гг. Коллективная монография под ред. А.М. Елизарова. Казань: Изд-во Казан. ун-та, 2008. С. 456-476.

15. *Елизаров А.М., Липачёв Е.К., Малахальцев М.А.* Веб-технологии для математика. Основы MathML. М.: Физматлит, 2010. 194 с.

16. *Елизаров А.М., Липачев Е.К., Малахальцев М.А.* Технологии Semantic Web в практике работы электронного журнала по математике // Труды 8 Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2006, Суздаль. Ярославль: Ярославский госуниверситет, 2006. С. 215-218.

17. *Ausbrooks R., Buswell S., Carlisle D., Chavchanidze G.* etc. Mathematical Markup Language (MathML) Version 3.0 2nd Edition. W3C Recommendation 10 April 2014 // World Wide Web Consortium (W3C). 2014. URL: http://www.w3.org/TR/ MathML/mathml.pdf.

18. *Kohlhase M., Şucan I.A.* A search engine for mathematical formulae // International Conference on Artificial Intelligence and Symbolic Computation. 2006.

19. *Muhammad Adeel, Hui Siu Cheung, Sikandar.* Math GO! prototype of a content based mathematical formula search engine // Journal of Theoretical and Applied Information Technology. 2008.

20. *Wiki Pages – web-xslt. Example XSLT code for transforming XML languages for the web.* URL: https://code.google.com/ p/web-xslt/.

21. *Kohlhase M.* MathML presenting and capturing mathematics for the Web. URL: http://www.w3.org/Math/Documents/ mathml-tutorial.pdf.

22. *SnuggleTeX – Overview & Features* [Электронный ресурс] // School of Physics and Astronomy. URL: http://www2.ph.ed.ac.uk/snuggletex/documentation/ overview-and-features.html.

23. *Working with MathML & Wolfram Language Documentation.* URL: http://reference. wolfram.com/language/XML/tutorial/MathML.html.

24. *MathJax. Beautiful math in all browsers.* URL: http://www.mathjax.org/.

25. *Getting Started – MathJax 2.5 documentation.* URL: http://docs.mathjax.org/en/latest/start.html.

26. *jQuery – write less, do more* [Электронный ресурс] URL: https://jquery.com.

SEMANTIC ANALYSIS OF DOCUMENTS IN THE CONTROL SYSTEM OF DIGITAL SCIENTIFIC COLLECTIONS

S.M. Khaydarov

Kazan (Volga region) Federal University
15jkeee@gmail.com

Abstract

Methods of the semantic documents parsing in digital control system of scientific collections, including electronic journals, offered. The methods of processing documents containing mathematical formulas and methods for the conversion of documents from the OpenXML-format in TeX-format considered. The search algorithm for the mathematical formulas in the collections of documents stored in OpenXML-format designed. The algorithm is implemented as online-service on platform science.tatarstan.

Keywords: semantic analysis, publishing systems.

REFERENCES

1. *Elizarov A.M., Lipachev E.K., Khokhlov Yu.E.* Semanticheskie metody strukturirovaniya matematicheskogo kontenta, obespechivayushchie rasshirennuyu poiskovuyu funktsional'nost' // Informatsionnoe obshchestvo. 2013. № 1–2. S. 83-92.

2. *Elizarov A.M., Lipachev E.K., Malakhaltsev M.A.* Veb-tehnologii v rabote elektronnogo matematicheskogo zhurnala Lobachevskii Journal of Mathematics // V sbornike: Nauchnyiy servis v seti Internet: mnogoyadernyiy kompyu-ternyiy mir. 15 let RFFI. Trudyi Vserossiyskoy nauchnoy konferentsii. Moskovskiy gosudarstvennyiy universitet im. M.V. Lomonosova, Yuzhnyiy federalnyiy universitet, Institut vyichislitelnoy matematiki RAN. g. Moskva, 2007. S. 355-356.

3. *Elizarov A.M., Zuev D.S., Lipachev E.K.* Informatsionnye sistemy upravleniya elektronnymi nauchnymi zhurnalami // Nauchno-tekhnicheskaya informatsiya. Seriya 1. Organizatsiya i metodika informatsionnoy raboty. 2014. № 3. S. 31-38.

4. *Khaydarov Sh.M.* Metody upravleniya matematicheskim kontentom v informatsionnykh izdatel'skikh sistemakh // Tr. Matem. tsentra im. N.I. Lobachevskogo. Materialy 14-y Vseros. Molodezhnoy shkoly-konferentsii «Lobachevskie chteniya–2015» (Kazan', 22–27 oktyabrya 2015 goda). Kazan'. 2015. S. 162-165.

5. *Vouter V.V.* Open XML – Kratko i dostupno. Open XML Technical Evangelist, Microsoft, 2007. 101 s.

6. *Standard ECMA-376: Office Open XML File Formats*. URL: http://www.ecmainternational.org/publications/standards/Ecma-376.htm

7. *Lipachev E.K., Khaydarov Sh.M.* Sistema servisov preobrazovaniya elektronnykh matematicheskikh dokumentov na osnove oblachnykh tekhnologiy // Trudy Matematicheskogo tsentra im. N.I. Lobachevskogo. Kazan'. 2013. T. 47. S. 109-110.

8. *Elizarov A.M., Lipachev E.K., Nevzorova O.A., Solov'ev V.D.* Metody i sredstva semanticheskogo strukturirovaniya elektronnykh matematicheskikh dokumentov // Doklady Akademii nauk. 2014. T. 457. № 6. S. 642-645.

9. *Akhmetov D.Yu., Gerasimov A.N., Grachev A.O., Elizarov A.M., Lipachev E.K.* Oblachnaya platforma podderzhki elektronnykh nauchnykh izdaniy // Uchenye zapiski Instituta sotsial'nykh i gumanitarnykh znaniy. 2014. № 1 (12), ch.1. S. 13-19.

10. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A., Solovyev V.D., Zhiltsov N.G.* Mathematical knowledge representation: semantic models and formalisms // Lobachevskii Journal of Mathematics. 2014. V. 35. No 4. P. 348-354.

11. *Biryal'tsev E.V., Elizarov A.M., Zhil'tsov N.G., Lipachev E.K., Nevzorova O.A., Solov'ev V.D.* Methods for analyzing semantic data of electronic collections in mathematics // Automatic Documentation and Mathematical Linguistics. 2014. V. 48, No 2. P. 81-85.

12. *Biryal'tsev E.V., Galimov M.R., Zhil'tsov N.G., Nevzorova O.A.* Podkhod k semanticheskomu poisku matematicheskikh vyrazheniy // OSTIS-2012. 2012. S. 245–256.

13. *Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E.* OntoMathPRO Ontology: A Linked Data Hub for Mathematics // Knowledge Engineering and the

Semantic Web Communications in Computer and Information Science. 2014. V. 468. P. 105-119. URL: http://link.springer.com/chapter/10.1007/978-3-319-11716-4_9.

14. *Veselago V.G., Elizarov A.M., Lipachev E.K., Malakhal'tsev M.A.* Formirovanie i podderzhka fiziko-matematicheskikh elektronnykh nauchnykh izdaniy: perekhod na tekhnologii Semanticheskogo Veba // Nauchno-issledovatel'skiy institut matematiki i mekhaniki im. N.G. Chebotareva Kazanskogo gosudarstvennogo universiteta. 2003–2007 gg. Kollektivnaya monografiya pod red. A.M. Elizarova. 2008. S. 456-476.

15. *Elizarov A.M., Lipachev E.K., Malakhal'tsev M.A.* Veb-tekhnologii dlya matematika. Osnovy MathML. Moskva: Fizmatlit, 2010. 194 s.

16. *Elizarov A.M., Lipachev E.K., Malakhaltsev M.A.* Tekhnologii Semantic Web v praktike raboty ehlektronnogo zhurnala po matematike // Trudy 8 Vserossijskoj nauchnoj konferentsii «Elektronnye biblioteki: perspektivnye metody i tekhnologii, ehlektronnye kollektsii" – RCDL'2006, Suzdal'. Yaroslavl': Yaroslavskij gosuniversitet, 2006. S. 215-218.

17. *Ausbrooks R., Buswell S., Carlisle D., Chavchanidze G. etc.* Mathematical Markup Language (MathML) Version 3.0 2nd Edition. W3C Recommendation 10 April 2014. // World Wide Web Consortium (W3C). 2014. URL: http://www.w3.org/TR/ MathML/ mathml.pdf

18. *Kohlhase M., Şucan I.A.* A Search Engine for Mathematical Formulae // Inter-national Conference on Artificial Intelligence and Symbolic Computation. 2006.

19. *Muhammad Adeel, Hui Siu Cheung, Sikandar*. Math GO! prototype of a content based mathematical formula search engine // Journal of Theoretical and AP.ied Information Technology. 2008.

20. *Wiki Pages – web-xslt. Example XSLT code for transforming XML languages for the web.* URL: https://code.google.com/p/web-xslt/.

21. *Kohlhase M.* MathML Presenting and Capturing Mathematics for the Web // World Wide Web Consortium (W3C). URL: http://www.w3.org/Math/Documents/ mathml-tutorial.pdf.

22. *SnuggleTeX – Overview & Features* // School of Phys-ics and Astronomy. URL: http://www2.ph.ed.ac.uk/snuggletex/documentation/overview-and-features. html.

23. *Working with MathML& Wolfram Language Documentation* // Wolfram: Computation Meets Knowledge. URL: http://reference.wolfram.com/language/XML/tutorial/MathML.html.

24. *MathJax. Beautiful math in all browsers*. URL: http://www.mathjax.org/.

25. *Getting Started – MathJax 2.5 documentation.* URL: http://docs.mathjax.org/en/latest/start.html

26. *jQuery – write less, do more.* URL: https://jquery.com.

СВЕДЕНИЯ ОБ АВТОРЕ

ХАЙДАРОВ Шамиль Махмутович – аспирант Института математики и механики им. Н.И. Лобачевского Казанского (Приволжского) федерального университета.

Shamil Mahmutovich KHAYDAROV, received MS degree in mathematics from Kazan Federal University (2015). Currently is a graduate student at the N.I. Lobachevskii Institute of Mathematics and Mechanics of Kazan Federal University. Current scientific interests: data mining, recommender systems, cloud computing, knowledge extraction technologies.

email: 15jkeee@gmail.com

Материал поступил в редакцию 12 февраля 2015 года