ИПМ им.М.В.Келдыша РАН

Абрау-2017 • Труды конференции

Е.И. Моисеев, А.А. Муромский,
Н.П. Тучкова

**Уточнение поискового запроса с
помощью символьной информации**

Рекомендуемая форма библиографической ссылки

Моисеев Е.И., Муромский А.А., Тучкова Н.П. Уточнение поискового запроса с помощью символьной информации // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18-23 сентября 2017 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2017. — С. 356-362. — URL: <http://keldysh.ru/abrau/2017/34.pdf> doi:[10.20948/abrau-2017-34](https://doi.org/10.20948/abrau-2017-34)

Размещена также [презентация к докладу](#)

Уточнение поискового запроса с помощью символьной информации

Е.И. Моисеев^{1,2}, А.А. Муромский², Н.П. Тучкова²

¹ ФОУ ВПО "Московский государственный университет имени
М.В.Ломоносова", факультет ВМК

²Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН

Аннотация. Исследование посвящено использованию символьных выражений в информационных запросах при поиске публикаций из математических предметных областей. Практика показывает, что подавляющее большинство зарубежных и отечественных журналов и издательств работает над созданием собственных стилевых файлов для использования TeX-нотации в оформлении статей. Несмотря на известное многообразие записей математических выражений их можно систематизировать благодаря контексту, традициям математических школ и стандартам, задаваемым стилевыми файлами. Накапливая формулы, соответствующие определенному контексту, можно дополнить ими статьи предметного тезауруса и использовать в дальнейшем эту символьную информацию в поисковом запросе.

Ключевые слова: сравнение математических текстов, семантический поиск, тезаурус

Рассмотрим пример синтаксической обработки текста статьи для формирования поискового образа по предметной области «волновое уравнение математической физики».

Пример 1.

Задача о малых поперечных колебаниях струны в плоскости

Струна: тонкая однородная гибкая одинаковой (постоянной) толщины, не оказывает сопротивления изгибу, нерастяжима ($\Delta x \approx 0$), т.е. длина струны l не изменяется, что при малых колебаниях допустимо ($\Delta x \ll l$). Вес и толщина струны не учитываются.

Однородное уравнение колебания струны:

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}, \quad -\infty < x < \infty, \quad t > 0, \quad a^2 - \text{константа.}$$

Начальные условия $\left\{ \begin{array}{l} u(x,0) = \varphi(x), \\ u'(x,0) = \psi(x) \end{array} \right\}$. Коши задача

Граничные условия $\left\{ \begin{array}{l} u(0,t) = 0, \\ u'(l,t) = 0 \end{array} \right\}$.

Общее решение: $u(x, t) = \varphi(x - at) + \psi(x + at)$, где φ и ψ - дважды дифференцируемые функции.

Даламбера формула: $u(x, t) = \frac{1}{2} [\varphi(x - at) + \varphi(x + at)] + \frac{1}{2a} \int_{x-at}^{x+at} \psi(z) dz$

Метод Фурье.

Задача: найти $u(x, t)$, такую, чтобы $-\infty < x < \infty$, $t \geq 0$, $\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}$,
 $u(x, 0) = f(x)$

Решение

$$u(x, t) = X(x)T(t), \quad \frac{T'(t)}{a^2 T(t)} = \frac{X''(x)}{X(x)}, \quad \begin{cases} T'(t) + \mu a^2 T(t) = 0 \\ X''(x) + \mu X(x) = 0 \end{cases}$$

$$u(x, t) = \frac{1}{2a\sqrt{\pi t}} \int_{-\infty}^{+\infty} f(\tau) e^{-\frac{(\tau-x)^2}{4a^2 t}} d\tau, [1]).$$

Выпишем ключевые слова, соответствующие данному фрагменту текста.

Ключевые слова (о колебании струны):

- малые плоские поперечные колебания струны,
- вертикальное перемещение струны,
- закон свободного колебания струны, Даламбера решение, Даламбера формула,
- общее решение однородного волнового уравнения,
- однородное волновое уравнение,
- уравнение малых поперечных колебаний струны, однородная среда колебаний,
- волновое уравнение с одной пространственной переменной, бегущие волны,
- среда колебаний без сопротивления,
- правая бегущая волна, левая бегущая волна,
- Даламбера метод, начальные и граничные условия,
- характеристики для одномерного однородного волнового уравнения,
- Коши задача для одномерного однородного волнового уравнения,
- метод разделения переменных, метод Фурье,
- основные уравнения математической физики,
- уравнение свободных колебаний струны,
- основные задачи математической физики,
- уравнение в частных производных второго порядка с постоянными
- коэффициентами.

Ключевые слова с количеством упоминаний и формулы составляют поисковый образ документа и входят в описание предметной области исследований автора (публикации). Аккумулируя статьи и их поисковые образы, устанавливаем связи между терминами и получаем вариант тезауруса для предметной области автора и соавторов. При добавлении библиографических записей из списков цитируемой литературы присоединяем разделы предметной или смежных областей. Такой способ разметки публикаций используется давно для реферативных журналов и классификаторов таких УДК[2] и MSC [3], однако недостаточно применяется в цифровых библиотеках, и с учетом символьной информации особенно. Тем не менее, есть случаи, когда использование формул для разметки публикаций вносит существенную дополнительную информацию, позволяет установить дублирование результатов и др.

Можно проследить, как менялись представления и математическая запись некоторых уравнений математической физики. Отличия связаны с развитием определенных тенденций в различных математических школах, регионах и странах. Это нередко сказывается на современных работах и часто приводит к взаимному недопониманию ученых, несмотря на исследования в одной предметной области.

Есть также исторические сложившиеся формулировки и определения, вызывающие затруднения при современном восприятии.

Сопоставление формул и сравнение различных математических записей для улучшения понимания сути, отслеживания дублирования, плагиата и установления приоритетов может быть полезно как в исследованиях, так и в методических целях при обучении.

Например, в рамках группы академика РАН Е.И. Моисеева разрабатываются тезаурусы по волновому уравнению, специальным функциям, градиентным теориям механики сплошных сред [4-7]. Все это может быть частью ресурса по предметной области "краевые задачи математической физики", представлено средствами LibMeta[8] и способствовать развитию как исследуемых предметных областей, так и информационных технологий в разделе поиска и управления информацией на основе тезаурусов предметных математических областей.

TeX-нотация и поиск

Обратимся к основным особенностям работы с математическими текстами, характерными для издательских систем и журналов. Это — работа с текстами в TeX-нотации [9]. За многие годы существования Тех-трансляторов популярность этой технологии форматирования текстов с математическими записями невероятно возросла, а многие издания такого характера вырабатывают свой стиль, который становится основой для оформления текстов и подачи публикаций в издательство. Это довольно трудоемкое занятие, требующее аккуратности и терпения, как и любой программистский

труд, стало обычным для научного сотрудника, поскольку результат форматирования известен, а правила журналов неоспоримы. Известно, что Дональд Кнут [9] при создании TeX-транслятора стремился к совершенству, поскольку, впервые столкнувшись с недостатками форматирования текстов статей, понял, что автоматизация этого процесса будет востребована и положительно воспринята научным сообществом.

Вот пример, рассмотренный ранее и воспроизведенный в варианте TeX-нотации.

Пример 2.

```
\documentclass{ccasmon}
\begin{document}
\title{Пример}
\keywords{волновое уравнение}
\section{Задача о малых поперечных колебаниях струны в плоскости}
```

Струна: тонкая однородная гибкая одинаковой (постоянной) толщины, не оказывает сопротивления изгибу, нерастяжима

$\Delta x \approx 0$, т.е. длина струны l не изменяется, что при малых колебаниях допустимо $\Delta x \ll 0$. Вес и толщина струны не учитывается.

Однородное уравнение колебания струны

$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}$, $-\infty < x < \infty$, $t > 0$, a^2 - константа.

Коши задача. Начальные условия

```
\left\{
\begin{array}[pos]{spalten}
u(x,0) = \varphi(x), \\
u'_t(x,0) = \psi(x)
\end{array}
\right\}
```

Граничные условия

```
\left\{ \begin{array}[pos]{spalten}
u(0,t) = 0, \\
u'(l,t) = 0
\end{array}
\right\}
```

Общее решение: $u(x,t) = \varphi(x-at) + \psi(x+at)$, где φ и ψ - дважды дифференцируемые функции.

Даламбера формула: $\frac{1}{2}[\varphi(x-at) + \varphi(x+at)] + \frac{1}{2a} \int_{x-at}^{x+at} \psi(z) dz$

Метод Фурье.

Найти $u(x,t)$, такую, чтобы $-\infty < x < \infty$, $t > 0$, $\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2}$ $u(x,0) = f(x)$.

Решение

$$u(x,t) = X(x) T(t), \quad \frac{T'(t)}{a^2 T(t)} = \frac{X'(x)}{X(x)},$$

$$T'(t) + \mu a^2 T(t) = 0, \quad X''(x) + \mu X(x) = 0$$

$$u(x,t) = \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{+\infty} f(\tau) e^{-\frac{(\tau - x)^2}{4a^2 t}} d\tau, \quad [4]$$

Понятно, что такой текст уже может быть подвергнут семантической обработке (что, собственно, и делается средствами трансляции) и проанализирован с точки зрения повторяющихся элементов.

На рис. 1 приведем схему поиска по формулам в общем виде.

Рис. 1. Схема поиска по формулам в общем виде

Здесь используются следующие сокращения:

1. Словесная формулировка задачи (СФЗ)
2. Представление задачи в виде совокупности ключевых слов(ПЗ)
(дескриптор, синонимы, ассоциативные термины)
3. Математическая запись задачи (МЗЗ)
4. Новая задача (НЗ)
5. Известная задача (ИЗ)
6. Поиск на новизну задачи (ПНЗ)
7. Известный метод решения (ИМРЗ)
8. Новый метод решения (НМРЗ)
9. Математическая запись решения (МЗРЗ)

На рис. 1 иллюстрируется технология поиска на основе вторичных документов с использованием символического представления данных в виде

формул. Эта технология также включает сравнение текстов по аналогии, механизм пополнения тезауруса, а также возможность установления связей "от формулы к тексту" и "обратной" связи "от текста к формуле". Последнее реализуется за счет "тезауруса формул" по предметной области, в котором символьное представление сопровождается текстовым (в виде названий, определений, ключевых слов), что позволяет сочетать сравнение ключевых слов и символьных выражений.

Заключение

Учитывая, что в сложившихся коллективах (школах) есть тенденция к сохранению обозначений, а также, что известные журналы имеют свои TeX-стили, можем использовать устоявшиеся обозначения (конструкции TeX-выражений) для дополнения терминов математических тезаурусов по определенной предметной области. Эта информация в дальнейшем может быть использована в поисковых образах по некоторой математической предметной области. Отметим, что происходят качественные изменения в развитии механизмов представления научных публикаций. Наблюдается усовершенствование систем ввода данных, например, появились и стали стандартом интерфейсы для размещения публикаций в базах данных конференций и журналов в том числе и на русском языке с использованием TeX-выражений. При визуализации и хранении данных используется MathML и другие средства. Продолжается развитие методов семантической обработки информации, включая использование тематических тезаурусов и онтологий предметных областей.

Перечисленное представляет возможности для нового подхода в организации поиска по символьной информации, и в частности, математическим формулам [10, 11]. Эти возможности задействованы при разработке библиотеки LibMeta. Для текстов LibMeta, имеющих представление в MathML реализуется сравнение, а для TeX-формул формируется соответствующее MathML-представление для размещения в базу.

В рамках текущего исследования предложены:

- способ сравнения текстов с использованием символьной информации;
- схема пополнения тезауруса;
- схема сравнения формул;
- технология поиска на основе вторичных документов с использованием символьного представления данных;
- схема уточнения запроса на основе символьной информации.

Проблему, безусловно, составляет PDF-представление публикаций в доступных источниках, базах данных издательств и журналов. Надо заметить, что простое заимствование данных из интернет-ресурсов в данной работе не предполагается. Наполнение библиотеки требует договорных отношений и соответствующих соглашений с такими (и другими, традиционными) ресурсами. В этом случае можно договариваться о цитировании внешних

ресурсов на уровне первоисточников, которые, как правило, есть у издателей в исходных вариантах (TeX или Word-файлах). Если предоставлен Word-файл, то для размещения в базе библиотеки используется текстовая часть статьи тезауруса, включающая определения и названия формул. Применяются механизмы сравнения пристатейных ключевых слов с терминами тезауруса, а также частотный анализ текста для выявления главных терминов с добавлением их в списки ключевых слов. Далее публикация индексируется теми формулами, которые имеются в тезаурусе в соответствующих ключевым словам статьях. Реализуется также "обратный" механизм: от текста к формуле. Понятно, что, вполне возможно, что в индексируемой работе могут не встречаться найденные по ключевым словам в тезаурусе формулы, но определение места в предметной области тезауруса работе при таком алгоритме гарантируется.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект № 17-07-00217а.

Литература

1. Тихонов А.Н., Самарский А.А. Уравнения математической физики. М.: Гостехтеоретиздат. 1953. 753 с.
2. Справочник по УДК. <https://teacode.com/online/udc/>.
3. Mathematical Subject Classification - MSC2010. <https://zbmath.org/classification/>
4. Lewis A.C. Kenneth O. May and information retrieval in mathematics // *Historia Mathematica* V. 31, Issue 2, May 2004, P. 186–195.
5. Моисеев Е.И., Муромский А.А., Тучкова Н.П. Интернет и математические знания: представление уравнений математической физики в информационно-поисковой среде. М.: МАКС Пресс, 2008. 80 с.
6. Моисеев Е.И., Муромский А.А., Тучкова Н.П. Тезаурус информационно-поисковый по предметной области «обыкновенные дифференциальные уравнения». М.: МАКС Пресс, 2005. 116 с.
7. Muromskiy A.A., Tuchkova N.P. Thesauruses and its application to the problem of situation awareness // *Proceedings & Program of the International Workshop "CI:CM/IACC/CS-2017"*, March 12-19, 2017. P. 38-40.
8. Серебряков В.А., Атаева О.М. Информационная модель открытой персональной семантической библиотеки LibMeta // Труды XVIII Всероссийской научной конференции "Научный сервис в сети интернет". Новороссийск, 19 по 24 сентября 2016 г. ИПМ им. М.В. Келдыша РАН. С. 304-313.
9. Кнут Д.Э. Всё про TeX. М.: Издательский дом «Вильямс». 2003 560 с.
10. Misutka J., Galambos L. Extending Full Text Search Engine for Mathematical Context. Towards Digital Mathematics Library. Birmingham, United Kingdom, July 27th, 2008. Masaryk University, Brno, 2008. Pp. 55-67.
11. Mathematics Stack Exchange URL: <http://math.stackexchange.com>.