ТРАНСЛЯЦИЯ МАТЕМАТИЧЕСКИХ ТЕКСТОВ В НОТАЦИИ TEX В ФОРМАТ, СОХРАНЯЮЩИЙ СЕМАТИНТИКУ

А.А. Кондратович

Белорусский государственный университет информатики и радиоэлектроники, Минск, Беларусь, andrew.kondratovich@gmail.com

Abstract. This article describes the problem of storing and extracting information from semistructured texts. It also describes the traditional ways and methods for representation of mathematical formulas and problems of this approach. The way of solving this problem in a particular case, using translation of source texts, is described too. This article can be applied to systems of distance learning systems, machine learning and intelligent systems.

Компьютерное и дистанционное обучение сегодня имеют высокие темпы развиятия и становятся все теснее связаны с областью искусственного интеллекта. В настоящее время проводятся активные исследования в области методов, систем и средств для интерактивного обучения математике [1], для тестирования по естественнонаучным дисциплинам [2], для обучения и подготовке по математическим и статистическим дисциплинам на безе "облачных" технологий [3] и др. смежных областях. Попытка обзора последних решений в области обучения математики предпринята в работе [4].

Особенностью многих научных и образовательных материалов, содержащих математику, физику, экономические науки и т.п., является наличие специфических естественнонаучных текстов – математических формул [5].

Исторически сформировалась стандартная нотация математических формул, использующаяся сегодня на бумажных носителях в виде разметки на плоскости букв разных алфавитов, специальных математических символов и символических выражений.

Для разметки текстов широко используется нотация TeX, разработанная еще Д. Кнутом [6]. Сегодня издается, по разным оценкам, от 20 до 30 тысяч электронных научных журналов, только IEEE ежегодно публикует более полумиллиона страниц технических работ. Все крупнейшие издательства публикуют свои журналы в электронном виде. К сожалению, при использовании TeX теряется информация о семантике математических выражений – содержание растворяется в представлении. Однако такой формат предоставления текстов настолько распространился, что уже практически нигде не принимаются ни для редактирования, ни даже для ознакомления рукописи в других форматах, например, в формате Mcrosoft Word, поскольку отражение его формул несовместимо с TeX. Во многих развитых компьютерных аналитических системах, например, Maple, Mathematica, Maxima возможен экспорт документов в формат *.tex. Для представления формул в системе MediaWiki также используется TeX-нотация.

С наступлением информационной эры появилась компьютеризированная нотация

математических формул. Сначала это был синтаксис арифметических выражений первых языков программирования (Fortran, Algol, Basic, Pascal), специализированных математических пакетов (Maxima, Maple, MathCAD), тестовых редакторов, а в дальнейшем были разработаны стандарты языков разметки текста (HTML, MathML, TeX и др.).

В случае представления формул в программировании их представление определяется последовательностью их обработки. Нотация языков программирования

Библиотека БГУИР

позволяет сохранить семантику математических формул. Она приспособленная для вычислений по формулам и является инструментом всех программистов, но не воспринимается людьми, далекими от программирования, не принимается журналами к печати и т.д.

Существующая ситуация такова, что наиболее распространенные системы хранение и представления данных представляют собой слабоструктурированные источники информации. Основными требованиями к данным системам является удобство форматирования и верстки, распространенность формата нотации, совместимость. Требование сохранения семантики текста в большинстве случаев отсутствует.

Интеллектуальные системы, системы машинного обучения, системы дистанционного обучения и самообучения оперируют семантикой текстов. Поэтому формат данных в этих системах далек от естественно-языкового представления, но сохраняет семантику и содержимое информации.

Для возможности использования всего богатства существующих информационных источников компьюерные системы должны научиться понимать слабоструктурированные тексты, близкие к естественно-языковым. В рамках проекта OSTIS[7] возникла задача автоматического перевода математических текстов из нотации LaTeX в формат представления данных в семантических сетях SCS. Эту задачу стоит понимать как проблему трансляции текстов исходного языка в целевой.

Трансляция представляет собой обработку набора символов в целях извлечения их значения и сохранения в другом виде[8]. Обычно, задачу распознавания делят на 2 стадии – лексический и синтаксический анализ.

В информатике лексический анализ — это процесс аналитического разбора входной последовательности символов (например, такой как последовательность символом математического выражния) с целью получения на выходе последовательности символов, называемых «токенами». Группа символов входной последовательности, идентифицируемая на выходе процесса как токен, называется лексемой. Лексемы обладают определенной смысловой нагрузкой.

Цель такой конвертации обычно состоит в том, чтобы подготовить входную последовательность для другой программы – синтаксического анализатора, и избавить его от определения лексических подробностей в контекстно-свободной грамматике.

Как правило, лексический анализ производится с точки зрения определённого формального языка или набора языков. Язык, а точнее его грамматика, задаёт определённый набор лексем, которые могут встретиться на входе процесса.

Синтаксический анализ представляет собой процесс сопоставления линейной последовательности лексем языка с его формальной грамматикой. Как правило, результатом синтаксического анализа является синтаксическая структура предложения, представленная либо в виде дерева зависимостей, либо в виде дерева составляющих, либо в виде некоторой комбинации первого и второго способов представления.

Для реализации транслятора математических выражений была выбран Pyparsing – это библиотека классов Python, которая позволяет быстро и легко создавать рекурсивно-нисходящие парсеры.

С помощью модуля pyparsing, сначала определяются базовые части грамматики. Затем они комбинируются в более сложные выражения для различных ветвей полного грамматического синтаксиса. Их комбинирование возможно с помощью определения связей, таких как:

–Какие выражения должны следовать друг за другом в грамматике.

Библиотека БГУИР

–Какие выражения являются заменами друг друга в определенном случае в грамматике.

–Какие выражения являются необязательными.

–Какие выражения являются повторяющимися.

Не смотря на то, что некоторые сложные грамматики могут иметь десятки или даже сотни грамматических комбинаций, много задач парсинга легко представляемы только с небольшим количеством определений. Представление грамматики в форме Бэкуса–Науэра помогает упорядочить структуру и дизайн транслятора. Она также помогает просмотреть путь прогресса и развития в реализации грамматики.

Работа транслятора включает в себя два шага. На первом шаге производится обработка и анализ исходных данных. Транслятор генерирует объектное представление математических формул в памяти. Результатом является граф зависимостей и отношений между элементами формулы. Данный этап позволяет совершить промежуточную обработку информации. Объеденяются одни и те же элементы, встречающиеся в разных местах. Происходит генерация промежуточных элементов формулы. Промежуточное представления исходных формул в виде графа является универсальных решением, которое позволит в случае необходимости расширить функционал и возможности программы.

Второй шаг представляет собой генерацию текстового проедставления формул на целевом языке. Он включает обход графа зависимостей и отношений между элементами формулы, полученного на предыдущем шаге, результатом которого является исходных текст на языке SCS.

Таким образом, транслятор математических выражений позволяет извлекать математические формулы из внешних источников и генерировать соответствующее представления на языке, понятном интеллектуальным системам проекта OSTIS.

Литература

1. Li Yang, Mo Qian, Wang Fang. An Interactive Mathematics Education Platform Based on Topic-Based Deep Search. Second International Workshop on Education Technology and Computer Science (ETCS), 6-7 March 2010, Vol: 2, p.p. 163 – 169.c.

2. Gu Yue-sheng, Zhu Jia-yi. Uploading Strategy of the Formula in the Web-Based Mathematics Testing System. International Conference on Computer Science and Software Engineering, 12-14 Dec. 2008, Vol: 5, p.p. 624 – 626.

3. Sousse T. Learning Math and Statistics on the Cloud, Towards an EC2-Based Google Docslike Portal for Teaching / Learning Collaboratively with R and Scilab. 2010 10th IEEE International Conference on Advanced Learning Technologies. July 05- 07 2010, pp. 752-753.

4. Chaamwe N. Integrating ICTs in the Teaching and Learning of Mathematics: An Overview. Second International Workshop on Education Technology and Computer Science (ETCS), 6-7 March 2010, Vol: 2, p.p. 397 – 400.

5. Вовк А.И., Гирнык Д.А. Язык общения математиков в Интернете. В кн.: New Information Technologies in Education for all: State of the art and Prospects (ITEA-2007), Kiev, Ukraine, IRTC, 21-23 November 2007, p.p. 96 – 103.

6. Donald E. Knuth, The TeXbook (Reading, Massachusetts: Addison-Wesley), 1984.

7. OSTIS [Электронный ресурс]. – Open Semantic Technology for Intelligent Systems. – Режим доступа : http://www.ostis.net/.

8. Альфред В. Ахо, Моника С. Лам, Рави Сети, Джеффри Д. Ульман. Компиляторы: принципы, технологии и инструментарий = Compilers: Principles, Techniques, and Tools — 2-е изд. — М.: Вильямс, 2008. — ISBN 978-5-8459-1349-4.

Библиотека БГУИР