МАТЕМАТИЧЕСКИЕ МОДЕЛИ И ВЫЧИСЛИТЕЛЬНЫЕ СТРУКТУРЫ
(Вычислительные системы)
2004 год Выпуск 173

УДК 519.68+004.428

КАНОНИЧЕСКИЙ ТЕКСТОВОЙ ФОРМАТ, RTF И МАТЕМАТИЧЕСКИЕ ФОРМУЛЫ

А.В.Манцивода, В.А.Липовченко

Основным типом информации, представленной в Интернете является структурированная (размеченная) текстовая информация. В первую очередь — это информация, находящаяся в формате языка разметки HTML. Язык HTML обладает универсальной структурой, позволяющей использовать его с одинаковым успехом для представления материалов в самых разнообразных сферах и предметных областях. Структура HTML очень проста и легко усваивается даже теми пользователями, которые имеют весьма невысокий уровень подготовки. Вкупе со значительной степенью толерантности браузеров, позволяющих работать даже с неряшливо подготовленными и ошибочными HTML-документами, это послужило одним из важнейших стимулов для невероятно быстрого распространения Интернета. Однако данные качества привели и к значительным проблемам, быстро накапливающимся в мировой информационной среде. Сегодня Интернет представляет собой огромную аморфную массу информационных ресурсов самого разного уровня, мало приспособленную для автоматической обработки и допускающую только примитивный поиск по ключевым словам. В последнее время стали предприниматься значительные усилия по решению данных проблем роста. В частности, очень серьезную и все более возрастающую роль в обмене символьной информацией в Интернете стал играть строгий и гибкий язык разметки XML [1]. Активно

развиваются проекты, связанные с созданием систем метаописаний ресурсов (например, [2]). Появилось понимание того, что для решения проблем развития Интернета вполне может быть использована математическая логика. В этом плане показателен проект Semantic Web консорциума W3 [3].

В рамках этого направления нашей исследовательской группой развивается комплексный проект, ориентированный на создание методов эффективного представления и обработки текстовой информации и иных ресурсов в распределенных информационных средах. Ядром проекта является система логической обработки знаний в Интернете на базе онтологий. Методологической основой для построения данной технологии служит семантическое программирование [4]. Подход, основанный на семантическом программировании, является весьма гибким и позволяет подключить целый спектор методов и технологий, ориентированных на обработку данных самых разных типов. Онтологический блок при этом служит управляющим ядром, обеспечивающим взаимодействие разнообразных сервисов, ориентированных на обработку данных и знаний самых разных типов. В рамках проекта нами разрабатывается комплекс соответствующих методов (см., например [5–10]).

Каноническое представление текстовой информации

В рамках данного проекта также реализуется концепция канонического представления текстовой информации. Использование «канонического» формата представления текстов на основе набора открытых стандартов позволяет решать большое количество проблем. Под каноническим форматом представления текстов мы понимаем такой формат, который

1) базируется на принципе разделения представления и содержания документа;

2) основан на разметке текстов логического уровня;

3) использует для представления документов открытые широко признанные стандарты и спецификации;

4) удобен для автоматической обработки, распространения и хранения информации;

5) допускает возможность конверсии текстов, представленных в каноническом формате, в другие распространенные форматы представления естественно-научной информации и обратно;

6) допускает трансляцию в различные полиграфические представления, как электронные, так и бумажные;

7) допускает богатую «интеллектуальную» обработку документа достаточно легкими с алгоритмической точки зрения программными средствами.

Базой для создания такого формата служит расширяемый язык разметки XML, являющийся открытым стандартом, широко используемым в Интернете. Различные компоненты текстовых документов представляются диалектами XML в разных пространствах имен. Канонический формат текстовых данных должен обязательно учитывать специфику представления текстовой информации в различных отраслях человеческих знаний. Особо интересной здесь является задача построения методов работы с естественно-научными текстами. В естественно-научные тексты, наряду с другими объектами (изображениями, таблицами), включаются формулы -- математические и химические. Человечеством накоплен огромный объем бесценных материалов естественно-научного характера, который представлен в самых разнообразных форматах, зачастую не приспособленных для публикации в Интернете и электронной обработки.

Технологический подход, реализуемый нами, состоит в комплексе методов, обеспечивающих следующие возможности:

- конверторы естественно-научных текстов в канонический формат,
- методы хранения и интеллектуальной обработки информации, представленной в каноническом формате,
- визуализацию текстов в каноническом формате,
- конверторы, переводящие тексты из канонического формата в другие широко распространенные форматы представления.

Форматы MS Word

В первую очередь речь, конечно, идет о текстах, представленных в формате TeX и его «отпрысков» (LaTeX, AMSTex).

Однако в последнее время все большую долю естественно-научных текстов занимают тексты, подготовленные с помощью редактора Microsoft Word. Как следствие, большое число документов естественно-научного характера хранятся в форматах, поддерживаемых этим текстовым процессором. Это стало возможным благодаря включению в MS Word специальных средств для создания формул (MS Equation Editor, MathType). Поэтому комплексная технология обработки естественно-научных текстов будет неполной без учета форматов представления текстов, предлагаемых данным текстовым редактором, поскольку подход, реализованный в MS Word, более доступен массовому пользователю, чем профессиональные системы верстки естественно-научных текстов. Это является ключевым конкурентным преимуществом данного редактора, объясняющим его широкое распространение.

С точки зрения использования в распределенных информационных средах Word-форматы обладают рядом негативных черт. Основной формат, используемый в редакторе, это формат DOC. Он держит пальму первенства по неудобству использования и обработки. Спецификации строения DOC-формата принципиально закрыты (что соответствует общей политике MS, ориентированной на полную зависимость пользователя от компании). Отрицательную роль играет и ориентация на полиграфическую разметку текста, что ставит серьезные барьеры для автоматической обработки. Другой наиболее популярный формат — RTF (Rich Text Format) [11] — в отличие от предыдущего, создавался компанией Microsoft как стандартный формат для обмена текстовыми документами и имеет открытую спецификацию. К сожалению, принципы построения, реализованные в данном формате, устарели. Основным недостатком формата RTF является опять же направленность на полиграфическое представление текста, в котором теряется логическая структура документа. Полиграфическая направленность чрезвычайно затрудняет реализацию «интеллектуальных» сервисов, работающих с документами. С другой стороны, этот формат допускает включение внешних объектов и иных способов расширения (в том числе средствами разметки не полиграфического, а логического уровня), которые активно используются самим редактором MS

Word. Поэтому, как правило, RTF-документы в расширенном формате содержат намного больше информации, чем позволяет базовая спецификация RTF.

Еще один недостаток Word-форматов — низкая пригодность для публикации информации в Интернете. Приходится конвертировать текст в HTML, причем на сегодняшний день MS Word при конвертировании в HTML-формат использует устаревший метод перевода формул в растровую графику (GIF-формат). Сами же тексты переводятся процессором MS Word в HTML-файлы катастрофически перегруженные специфической «технологической» информацией редактора, которая с точки зрения стандартной интернет-идеологии представляет собой мусор, который в значительной степени увеличивает трафик. Есть и другие довольно существенные недостатки. Вообще, с точки зрения современного понимания того, как работать со структурированной текстовой информацией, идеология, заложенная в редакторе Word, является более устаревшей, чем в более ранней системе LaTeX.

К настоящему времени созданы все необходимые технологические условия для эффективной работы даже с такими «тяжелыми» объектами, как Word-файлы. Созданы нужные открытые спецификации как для представления текстов произвольного характера, так и для представления таких специфических объектов, как математические формулы. Очень существенным шагом для работы с научными документами стало появление диалекта XML — языка MathML [12], ориентированного на описание математических формул. MathML интересен по многим причинам. Во-первых, формат MathML основан на языке разметки XML что обеспечивает стандартизированность структуры документов и позволяет использовать большое количество наработанных в мире компонент и библиотек. Во-вторых, формат MathML реализует два вида представлений: презентационное и содержательное. В-третьих, формат MathML является открытым стандартом, поддерживаемым консорциумом W3 [13], что очень важно для построения сложных систем, ориентированных на использование широкими массами пользователей.

В данной публикации мы имеем возможность анонсировать создание первой версии пакета инструментов для работы с естественно-научными текстами, созданными в редакторе MS Word. Пакет содержит

1. Парсер RTF-файлов.

2. Комплект методов, представленных в виде Java-библиотеки, а также интегрированных в функционально-логический язык Флэнг [14], который позволит реализовать новый подход в работе с RTF-документами, или точнее с их объектной моделью. Данные методы, в первую очередь, ориентированы на трансляцию RTF-документов в XML-формат, но также позволяют производить другие типы обработки RTF-документов.

3. Модуль, транслирующий формулы, представленные в закрытых форматах Equation Editor и MathType, в открытый формат MathML. Этот модуль позволяет работать с математическими формулами на уровне структуры, а не на уровне графических изображений, как в некоторых других RTF-конверторах.

4. Online-сервис по трансляции математических документов формата RTF в формат XML с трансляцией математических формул в формат MathML.

Данный пакет разработан в среде Java, что обеспечивает высокий уровень переносимости и технологичности. На базе инструментов пакета может строиться целый спектр разнообразных сервисов, работающих с естественно-научными текстами — как на локальном компьютере пользователя, так и на сервирах в online-режиме.

Объектная модель RTF

Рассмотрим более подробно объектную модель RTF-документов. Несмотря на сложность своей структуры, RTF-формат выстраивается в довольно простую объектную модель, состоящую всего из четырех основных элементов:

1) управляющее слово (ControlWord);

2) управляющий символ (ControlSymbol);

3) группа элементов (Group);

4) текст (Text).

Любой RTF-документ состоит преимущественно из команд управления, предназначенных для настройки программы чтения файлов (управляющие слова и управляющие символы).

Не вдаваясь в подробности RTF-формата, дадим краткое описание элементов структуры в контексте объектной модели.

Управляющее слово — основная управляющая команда в RTF. Элемент объектной модели ControlWord имеет два поля: word — непосредственно команда, param — параметр команды. Параметр может быть пустым.

Элемент ControlSymbol не имеет параметра и содержит единственное поле — symbol, определяющее непосредственно сам символ. Например, символ «~» задает жесткий (неразрываемый пробел) в RTF-формате.

Группой элементов будем называть набор элементов объектной модели, обособленный в элементе Group. Группы применяются для описания отдельных частей документа. Например, в группу объединяются команды для описания сносок, колонтитулов, абзацев и т.п. Также группы могут применяться для описания палитры цветов и набора шрифтов, используемых в документе.

В элементе объектной модели Text содержится единственное поле text, в котором могут содержаться названия шрифтов, описания объектов (формулы, графические изображения и т.д.) и непосредственно текстовая составляющая документа.

За трансляцию RTF-документов в объектную модель, с которой в дальнейшем работает пользователь, отвечает RTF-парсер.

Дальнейшая обработка документов производится с использованием специализированных методов объектной модели. Также RTF-парсер позволяет представлять объектную модель RTF в XML-формате, что дает возможность производить дальнейшую обработку с использованием методов, предназначенных для работы с XML.

Как уже было отмечено выше, нами разработана Java-библиотека, реализующая объектную модель RTF-документов. Эта библиотека использована для включения средств обработки текстов в функционально-логический язык Флэнг, ориентированный на обработку структурированной символьной

информации. Фленг является одним из ключевых элементов разрабатываемого нами подхода. С одной стороны, Флэнг — полноценный универсальный язык, включающий, в частности, все основные механизмы логического программирования. С другой стороны, богатые возможности работы с рекурсивными и древовидными структурами (термами), присущими логическому программированию, во Флэнге в максимальной степени "настроены" на обработку таких структур, как XML–, HTML–, и RTF–документы. Фленг также удобен для работы на "серверной" стороне, например, для автоматического генерирования HTML–страниц, создания сервлетов и т.д. Такие возможности Флэнга позволяют осуществить поддержку RTF–формата в различных сервисах.

Флэнг позволяет пользователю создавать собственные приложения и сервисы с поддержкой RTF–формата с «нуля», имея доступ к объектной модели RTF–документов. Но кроме этого на Фленге уже реализован конвертор из RTF– в XML–формат, который можно использовать как шаблон для создания собственных приложений.

Использование функционально–логического подхода для обработки текстовой информации в распределенных информационных сетях представляется весьма перспективным. В частности, применение функционально–логических средств в среде Java при создании продвинутых методов обработки текстовой информации позволяет в значительной степени сократить объем работы и повысить эффективность. Со временем данная технология будет развиваться, будут появляться новые шаблоны, упрощающие создание собственных приложения. Флэнг будет наполняться встроенными функциями, позволяющими автоматизировать многие процессы обработки текстовой информации естественно–научного характера. Например, встроенные функции будут позволять производить конвертирование из RTF в XML с различными уровнями детализации, т.е. достаточно будет указать значение параметра уровня детализации при конвертировании и на выходе получить соответствующей детализации XML–документ, что даст возможность производить конвертирование, отбрасывая вообще или в соответствующей мере стилевую информацию

из RTF-документов, сохраняя при этом базовую структуру документа.

Нужно отметить, что все вышесказанное в полной мере относится к работе с объектами, внедренными в RTF-документы к которым относятся графические изображения, формулы и т.п Доступ к таким объектам получить действительно несложно. Например, для того, чтобы получить доступ к информации, описывающей формулу, созданную в редакторе Equation Editor 3.0 нужно найти группу элементов с управляющим словом object которая содержит вложенную группу с управляющим словом objclass и текстом Equation 3. Следующая вложенная группа с управляющим словом objdata и будет содержать описание объекта.

Математические формулы

Обработка естественно-научных текстов, представленных в Word-формате невозможна без создания специальных средств обработки математических формул. Поэтому представляемый в данной публикации пакет включает разработанный нами модуль, позволяющий работать с формулами в форматах Equation Editor и MathType на уровне структуры, а не на уровне графических изображений.

Объекты Microsoft Equation, включаемые в RTF-документ для представления математических формул, состоят из графического изображения, дополняемого описанием структуры формулы, организованного в виде набора шестнадцатеричного кода

Проведенный нами анализ позволил установить механизмы описания структуры математических формул в данном формате. На основе данного анализа был разработан парсер объектов типа MS Equation. Кроме того, была реализована библиотека функций по работе с данным форматом и конвертированию формул из данного формата в формат MathML. Эта библиотека была включена в пакет обработки естественно-научных текстов, представленных в формате RTF. Апробирование и тестирование данного модуля показало высокую стабильность и эффективность его работы.

В настоящее время нами реализуется online-сервис, который получая на входе RTF-документ, автоматически конвертирует

его в документ в XML-формате. При этом сам текст документа транслируется в диалект DocBook [15], а математические формулы — в формат MathML. Кроме того, благодаря другим средствам, разрабатываемым в нашей группе, имеется возможность трансляции данного текста в формате (X)HTML с представлением математических формул как в формате MathML, так и в виде графических объектов. Нами планируется в ближайшее время открыть данный сервис в режиме свободного доступа.

Литература

1. Extensible Markup Language (XML) 1.0 (Third Edition). W3C Recommendation 04 February 2004. http://www.w3/org/TR/2004/REC-xml-20040204

2. Dublin Core Metadata Initiative (Дублинское ядро: инициативная группа по метаописаниям).http://dublincore.org

3. Semantic Web. http://www.w3.org/2001/sw/

4. GONCHAROV S.S., ERSHOV Yu.L., SVIRIDENKO D.I. Semantic programming// Information processing, Proc.IFIP 10-th World Comput.Congress, Dublin. – 1986. – P.1093–1100.

5. МАЛЫХ А.А., МАНЦИВОДА А.В. Система МЕТА и открытые модели знаний// Тр.Всероссийской науч.конф. «Научный сервис в сети Интернет-2004», Абрау-Дюрсо, 2004. – М.: Изд. МГУ, 2004. – С.173–175. Электронная версия http:// www.teacode.com/public/abrau-2004-2.txt

6. МАЛЫХ А.А., МАНЦИВОДА А.В. МЕТА: метаописания и образовательные пакеты // Тр.Всероссийской конф. Телематика'2004, С.-Петербург,2004. – С.-Петербург: Изд. ИТМО, 2004. – С.552–553.

7. МАНЦИВОДА А.В., ПЕТУХИН В.А. Порталы, обработка структурированной информации и языки искусственного интеллекта // Тр.Всероссийской конф. Телематика'2003. – С.-Петербург.2003. – С.168–169.

8. Математические формулы и электронные образовательные ресурсы /Манцивода А.В., Липовченко В.А., Малых А.А. и др. // Тр. Международного форума «Новые инфокоммуника

ционные технологии: достижения, проблемы, перспективы» . – Новосибирск, 2003. – С.78–84.

9. ЛИПОВЧЕНКО В.А., МАНЦИВОДА А.В. Трансляция математических формул из документов MS WORD в стандартный формат // Тр. Всероссийской конф. Телематика'2004. – С.-Петербург, 2004. – С.107–108.

10. КУРОПТЕВ А.А., МАНЦИВОДА А.В. Изображение математических формул в формате MathML //Тр.Всероссийской конф. Телематика'2004. – С.–Петербург, 2004. – С.110.

11. Rich Text Formate (RTF) Specification, version 1.6. http://msdn.microsoft.com/library/

12. MathML 2.0, a W3C Recommendation. http://www.w3.org/Math/

13. World Wide Web Consortium. http://www.w3.org

14. MANTSIVODA A. Flang: A Functional-Logic Language // Lecture Notes in Computer Science., 567, Processing Declarative Knowledge (eds. H.Boley and M.M.Richter)/ Springer. -1992. P.257-270.

15. DocBook Specification. http://docbook.org/

Поступила в редакцию
24 января 2005 года