ЭЛЕКТРОННЫЕ ИЗДАНИЯ И ПРЕДСТАВЛЕНИЕ МАТЕМАТИЧЕСКИХ ТЕКСТОВ НА WWW*

О. В. Олейник, Е. М. Толкачева, А. М. Федотов
Институт вычислительных технологий СО РАН
Новосибирск, Россия
e-mail: oleinik@adm.ict.nsc.ru, katerina@adm.ict.nsc.ru,
fedotov@adm.ict.nsc.ru

The problems of the automatization of the presentation of mathematical texts prepared in TEX system in the form of the WWW electronic publications. The existing approaches to the presentation of mathematical texts are analyzed and TEX-HTML converters are surveyed, their advantages and drawbacks are discussed. Technology is suggested based on the loading of mathematical fonts to the customer's computer and its implementation on the SB RAS server is described.

Осознание мировым сообществом стратегической роли информации стимулировало разработки новых информационных технологий как для получения и переработки больших объемов информации, так и для ее хранения и предоставления пользователям. В нашей стране особенно велико их значение в сфере науки и образования в связи с существующим информационным голодом, вызванным катастрофическим падением в последнее время объема подписки научно-технических библиотек, существенным сокращением числа получаемых библиотеками иностранных и отечественных изданий, а также падением тиражей и числа выходящих в свет русскоязычных книг и журналов. Известно, что не только развитие, но и поддержка научных исследований на должном уровне не мыслима без обмена информацией. Единственным выходом из создавшегося положения является использование сетевых информационных ресурсов мирового научного сообщества, предоставляемых Internet, и распространение своих достижений в виде электронных публикаций. В настоящей статье речь пойдет о технологии подготовки математических публикаций с использованием WWW для распространения в среде Internet.

1. Издательская система TEX

Постоянно совершенствующаяся технология подготовки оригиналов-макетов научно-технических текстов вызвала к жизни такие мощные текстовые процессоры, как TEX, MS Word, PageMaker и другие, им подобные, которые позволяют подготовить для печати

*Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований, грант №97-07-90372.
© О. В. Олейник, Е. М. Толкачева, А. М. Федотов, 1997.

практически любой документ. Однако, если речь идет о текстах, содержащих математические символы, то пока непревзойденной является издательская система TEX, созданная Д. Кнутом. В конце 70-ых гг. известный американский программист и автор знаменитого многотомного издания "Искусство программирования для ЭВМ"Дональд Кнут разработал программу для компьютерной верстки документации — TEX. В отличие от издательских систем, таких как Ventura и PageMaker, TEX — это свободно распространяемый продукт, что также способствовало росту его распространения. За более чем пятнадцатилетнюю историю своего существования TEX завоевал научный мир Америки, Западной Европы и нашей страны. Сейчас это стандарт *de facto* для подготовки научных публикаций, содержащих математические формулы.

Текстовый процессор TEX активно используются физиками, химиками, математиками и учеными других специальностей всего мира для обмена информацией и издательской деятельности. Причина популярности издательской системы TEX и ее применения для представления научно-технической информации заключается в высокой компактности, читаемости файлов вне TEXa, сохранении в них логической структуры документа и полной переносимости системы TEX на любые платформы. Электронная документация, подготовленная в TEXe, может быть воспроизведена практически в любых условиях — от PC до суперкомпьютера независимо от используемой операционной системы.

В связи с развитием гипертекстовых технологий Internet и все бóльшим применением последних для представления научно-технических текстов, на второе место выходит и другое преимущество издательской системы TEX, которое заключается в том, что по структуре команд язык программирования текстов TEX очень похож на язык программирования гипертекстов HTML (HyperText Markup Language) — язык гипертекстовой разметки документов, используемый для подготовки документов для WWW. Это свойство языка TEX, при наличии хорошо работающих конверторов или препроцессора, позволяет практически одновременно подготавливать электронную и печатную версии документов.

Вместе с компьютерным набором возникли две проблемы, связанные с использованием символов, не содержащихся в стандартной кодовой таблице (Code Page) символов US-ASCII (ISO-ASCII): первая — это проблема включения в подготавливаемые тексты символов национальных алфавитов и акцентированных символов, даже в странах, использующих в качестве основного латинский алфавит; вторая — проблема использования специальных символов (например, символов математических формул) и их представления на WWW.

2. Математика на WWW

Гипертекстовая разметка статьи, как правило, значительно облегчает чтение публикации с монитора, однако для математических текстов пока не существует более или менее пригодных для этого программных средств. Несмотря то, что язык HTML 3.2 имеет в своем арсенале поддержку некоторых математических формул, наиболее распространенные просмотрщики (WWW-браузеры), такие как Netscape Navigator или MS Internet Explorer, пока не поддерживают вывод математических символов и выражений. Это связано с тем, что в настоящее время отсутствуют какие-либо соглашения об унификации математических шрифтов на компьютере пользователя (точно так же, как отсутствуют соглашения об использовании дополнительных шрифтов на WWW-страницах).

Существуют специальные просмотрщики для математических текстов, такие как, на-

пример, Mathbrowser (http://www.mathsoft.com/browser/), рассчитанные на подготовку текста, с использованием специального для этого просмотрщика языка подготовки документов, отличного от HTML. Просмотрщик Mathbrowser в качестве математического формата использует формат документов Mathcad.

Универсальные просмотрщики, которые поддерживают стандарт математических формул языка HTML 3.0, как правило, являются коммерческими и не имеют широкого распространения в мире. Несмотря на их универсальность, возможности и интерфейс этих просмотрщиков пока оставляют желать лучшего (см., например, просмотрщики INRIA Amaya или MMM http://www.inria.fr/serveurs-eng.html). Кроме того, следует отметить, что долго обсуждавшийся стандарт языка HTML 3.0 (в котором предлагалась поддержка математики) так и не был принят сетевым сообществом и вместо него в декабре 1996 года узаконен стандарт HTML 3.2.

Поэтому подготовка математических текстов для WWW пока ведется в расчете на универсальные просмотрщики, имеющие широкое распространение, такие как Netscape Navigator или MS Internet Explorer, с использованием графических файлов для представления математических формул. Графические файлы для отображения математических формул можно представить пользователю двумя способами: 1) для каждой формулы иметь свою картинку для ее отображения, 2) используя Java приложения, загрузить на компьютер пользователя графические шрифты с математическими символами, из которых потом формировать математические выражения.

Первый подход реализован, например, в конверторе LaTeX2HTML (автор Nikos Drakos, University of Leeds, Великобритания, http://cbl.leeds.ac.uk/nikos/tex2html/tex2html.html). Недостатком этого подхода является то, что при подготовке документа вам неизвестна разрешимость экрана и размер используемых при просмотре шрифтов, установленных на компьютере пользователя, что приводит к несоответствию размеров текста и формул в документе, а это резко снижает его восприятие. С другой стороны, в документе, содержащем достаточно много формул, размер графических файлов становится весьма значительным, что замедляет его передачу по сети.

Подход, основанный на использовании Java-приложений и загрузке графических шрифтов, используется пока не очень широко для представления математических текстов. Его использование оправдано для представления текстов, содержащих большое количество математических формул. При этом подходе число передаваемых по сети графических файлов с математическими символами не зависит от величины текста, а при работе с большим количеством документов графические файлы будут браться из кэша, что существенно уменьшает нагрузку на сеть. Недостатком этого подхода является то, что на слабом компьютере Java-приложения работают достаточно медленно.

В качестве примера использования данного подхода можно привести пакет WEBeq, разработанный Геометрическим центром университета Миннесоты (Center for the Computation and Visualization of Geometric Structures, a National Science Foundation Science and Technology Center at the University of Minnesota). Данный подход реализован на WWW-сервере Сибирского отделения РАН (http://www- sbras.nsc.ru/) и доступен для свободного использования всеми, кто имеет IP соединение с “Сетью Internet Новосибирского научного центра"(http://www-sbras.nsc.ru/win/nsc-net/nsc.html). Подробности и инструкции на русском языке по использованию WEBeq можно найти в электронной публикации “Набор математических формул для WEB"(http://www.ict.nsc.ru/win/fedotov/web-eq/).

3. Обзор конверторов из TEX в HTML

В настоящее время многие солидные издательства выпускают печатную продукцию с помощью системы TEX, созданной Д. Кнутом и зарекомендовавшей себя как удобный инструмент для подготовки высококачественных печатных материалов. Однако, к сожалению, не удобно читать полученные по сети математические статьи, подготовленные, например, в формате LATEX или PlainTEX, не имея подходящего просмотрщика. Конечно, читатель может обработать данную статью транслятором TEX, получить DVI-файл и просмотреть статью с использованием какого-либо DVI-просмотрщика. Но здесь он сразу столкнется с двумя проблемами: первая — он должен установить на своей машине TEX и TEXовские шрифты, вторая — проблема совместимости (особенно трудная для русскоязычных текстов[1]), так как в статье могут быть использованы специальные форматы и стилевые файлы TEXa.

Поэтому наиболее правильным решением для читателя было бы конвертирование TEXовского файла в гипертекстовый документ. Как мы уже отмечали, попытка решить эту нетривиальную проблему впервые была предпринята в 1993 г. Н. Дракосом. Однако пока еще не существует полного решения проблемы преобразования математического текста, подготовленного в TEXe, в гипертекстовый документ.

Ниже перечислены известные авторам конверторы документов, подготовленных на основе издательской системы TEX в HTML. Ни один из существующих конверторов не работает полностью автоматически, особенно в части гипертекстовой структуры документа. Как правило, результирующий документ требует либо дополнительной правки, либо использования специальных стилей TEX, предназначенных для создания гипертекстовых документов.

LaTeX2HTML. LATEX 2.09 конвертор в HTML. Автор Nikos Drakos (University of Leeds, Великобритания). Конвертор написан на языке интерпретатора Perl (Perl scripts, объемом около 200 килобайт) для OS UNIX. Это наиболее полный из всех в настоящий момент работающих конверторов. Он включает математические формулы и таблицы в виде графических файлов, обрабатывает примечания и библиографию. Этот конвертор является, пожалуй, самым известным и хорошо зарекомендовавшим себя конвертором такого типа. В возможностях преобразователя присутствует разбор и раскрытие функций, описанных пользователем (`\newcommand`). Автоматически происходит преобразование LATEX-конструкций, подобных существующим в HTML: таблиц, подстрочных примечаний, списков и прочих. Не имеющие аналогов в HTML, элементы выделяются в отдельные файлы, компилируются с помощью LATEXa, а затем переводятся в формат Compuserve GIF и помещаются конвертором на соответствующие позиции внутри документа. Этот метод, конечно, привлекателен, однако, к сожалению, не лишен недостатков. Во-первых, он "медленный", поскольку требует слишком большого количества операций, а во-вторых, результат обработки не всегда корректен, и ошибки приходится править в ручную. Для работы требуются траслятор TEXa (формат LATEX 2.09), преобразователь DVI-файлов в PostScript (dvips), GhostScript и PBMPlus Toolkit. Все требуемое программное обеспечение и сам конвертор имеют статус freeware.

LaTeX2hyp. Программа, написанная на языке C, для трансляции документов из LATEXa в HTML, Text, TurboVision help, RTF или WinHelp RTF. Конвертор поддерживает перекрестные ссылки, библиографию, нумерацию и т.п. Полностью поддерживает

[1]См. статью А. В. Дорофеева и А. М. Федотова "Электронные публикации в среде Internet и множественность кодировок русского языка"(с. 31)

разметку текста LaTeX 2e в нотации HTML 3.0, ссылки и библиографию. Автор Roger Nelson, Вашингтонский университет.

Для трансляции таблиц и математических формул используется пакет math2html.

math2html. Транслятор математических формул и таблиц LaTeX в HTML 3.0, работающий под OS UNIX. Результат работы поддерживается коммерческим просмотрщиком Arena. Автор Йан Саарела (Janne Saarela e-mail: Janne.Saarela@hut.fi). Программа написана на C++ для любой UNIX-платформы, на которой были бы доступны flex, bison и (g)make. Производит перевод математических конструкций из LaTeXa 2.09 в HTML3. Необходимо отметить, что HTML3 обладает расширенными, по сравнению с HTML2, возможностями для работы с таблицами, формулами и уравнениями, поддержка именно этих возможностей и была реализована в math2html. Следует знать, что все математические формулы, включающие в себя структуры, не поддерживаемые math2html, будут проигнорированы. Единственным выходом из этого положения является перевод их в абсолютный код и отображение в конфигурационном файле.

vulcanize. Конвертор для OS UNIX, написанный на языке интерпретатора Perl (Perl scripts), для нематематических текстов, подготовленных LaTeX 2.09. Автор Mark-Jason Dominus (e-mail mjd@central.cis.upenn.edu).

Hyperlatex. Конвертор с подмножества LaTeX 2.09, работающий под OS UNIX. Предназначен для создания гипертекстовых страниц средствами LaTeX. Автор Otfried Schwarzkopf. Исходный текст написан на GNU Emacs Lisp для UNIX. Любители Emacs могут использовать конвертор прямо из него. Однако, если пользователь испытывает неприязнь к Emacs, он все равно сможет воспользоваться Hyperlatex'ом, запустив его из shell'a. Автор не преследовал цели поддерживать все команды LaTeXa, поэтому конвертор распознает весьма ограниченный набор команд. С непонятными ему конструкциями Hyperlatex борется, выдавая сообщения об ошибках. В случаях, когда необходимо отобразить, скажем, формулы или таблицы, что совсем не тривиальная задача, авторам представляется самым разумным перевести их в графический формат, а затем, поскольку существует возможность описывать те части текста, которые будут видны при просмотре стандартным WWW-браузером, отдельно от тех, которые пользователь предпочитает видеть в LaTeX-документе, в LaTeX-версии описать объект в LaTeX-формате, а в документ HTML поместить на нужном месте графическую вставку.

Если пользователь хочет отобразить i-й элемент какого-либо множества как в LaTeXe (n_i) и как в HTML-документе $n[i]$, то авторы этого преобразователя предоставляют ему такую возможность. Допускается также использование функций и командных скобок (режимов `\newenvironment`), определенных пользователем, но начинаться такие описания должны с новой строки, с команды `\H` или с пробела.

Tex2rtf. Конвертор LaTeX 2.09 в HTML, RTF, Windows Help RTF and wxHelp. Программа, написанная на C, работает на всех платформах (DOS, Windows, UNIX). Конвертор не поддерживает математику и имеет проблемы с переводом таблиц. Конвертор является частью свободно распространяемой библиотеки wxWindows, работающей под Sun Open Look, Motif, Windows 3.1, Windows 95/NT, non-GUI UNIX. При разборе текста пропускает таблицы и математику. Автор Юлиан Смарт (J.Smart@ed.ac.uk).

JAM. Мета-язык для описаний конверторов. Конвертирует в HTML текстовые файлы LaTeX и RTF. Конвертор написан на языке C для OS UNIX, Macintosh и Intel. Автор Lachlan Cranswick (e-mail lachlan@dmp.csiro.au).

BETA. Пакет, работающий под DOS. Конвертирует LaTeX в HTML. Организован в виде

формата TEX. Работает по аналогии с конвертором LaTeX2HTML. Автор Horst Wassenberg.

YODL. Язык создания документов с транслятором LATEX — HTML. В системе организована поддержка и некоторых других форматов. Понимает макросы TEXa. Правильно транслирует только текстовые документы. Работает под LINUX'ом (в принципе, может работать на любой UNIX платформе). Автор Karel Kubat (e-mail: karel@icce.rug.nl).

axTeX. Конвертирует текстовую часть LATEX-файла в HTML. Неконвертируемые части (формулы, таблицы и др.) вставляет в виде HTML-комментария в TEXовской нотации для последующей замены графическими файлами. Выделяет LATEX-объекты, встроенные в HTML-документ, создает отдельные .tex-файлы и, в конце, преобразовав их в графический формат, создает HTML-документ с соответствующими графическими вставками. Автор Philip Thrift.

HyperTeX. Наиболее полный интегрированный в WWW пакет, позволяющий конвертировать TEXовские документы, используя DVI или PDF-файлы. Автор Arthur Smith.

В сентябре 1994 г. Center for Geometry Analysis Numerics and Graphics заявил о появлении HyperTEX. На первом этапе была реализована версия для UNIX/X Window System, со временем был предложен вариант для Apple Macintosh. Для правильной работы требуется Acroexchange, желательно как можно более поздней версии. Исходные тексты написаны на C.

Вдохновленные успехом WWW, авторы программы расширили TEX возможностями добавления в документ связей между его частями и другими документами. Во время просмотра в режиме on-line связи становятся активными и дают пользователю возможность без труда передвигаться между документами и внутри них. Естественно, междокументные связи эффективны только на экране. Просмотр требует поддержки форматов файлов (.dvi, .ps и .pdf).

plain2. Автор Akihiro Uchida, Япония (e-mail uchida@ccs.mt.nec.co.jp). Стандартную поставку исходных файлов можно откомпилировать под System V или 4.xBSD UNIX, а также и MS DOS; для этого потребуется соответствующий компилятор C. Преобразует plain\text в форматы ROFF, HTML, TEX, LATEX.

plain2html. Написан программистами Colos (COnceptual Learning Of Science) Lyon, Франция. Язык, разработанный для этого конвертора, настолько похож на TEX, что может считаться его диалектом.

Таким образом, для разрешения проблемы успешного перевода документов из формата LATEX в формат HTML реально существуют две основные стратегии.

1. Поскольку HTML на сегодняшний день не имеет в своем арсенале возможностей поддержки математических формул в требуемом объеме, формулы предлагается превращать в изображения, а затем помещать на нужные места в документе. У этого довольно изящного метода, наиболее полно реализованного Н. Дракосом, есть ряд существенных недостатков: пользователь сталкивается с трудностями в согласовании шрифта, используемого для просмотра с размером графической вставки. Графическое изображение имеет фиксированное разрешение, поэтому, когда пользователь просматривает его на мониторе с разрешением 144 dpi, то рисунок выглядит вчетверо меньше, а при печати на принтере разрешением 600 dpi пикселы превратятся в квадраты, и шрифт формулы будет выглядеть крайне неаккуратно, несмотря на высокое качество устройства.

2. Создать новый язык и его транслятор и в дальнейшем использовать его для составления документов и последующего перевода в различные форматы, такие как LATEX, HTML, PS и др. Примеров несколько: JAM, BETA-format, YODL, HyperTeX, plain2html.

Стратегия обычно такова. Создаются новый язык и его транслятор, который в дальнейшем используется для составления документов и последующего перевода в различные форматы, такие как LaTeX, HTML, PS и др. (например, JAM, BETA Euromath System, YODL, HyperTeX, plain2html). Процесс конвертирования происходит по следующей схеме:

конвертируются конструкции TeX, имеющие аналоги в HTML, такие как таблицы, подстрочные примечания, списки и др.;

конструкции, которым не удается сопоставить адекватной замены, выделяются в отдельные файлы;

эти файлы обрабатываются транслятором LaTeX, и затем переводятся в графический формат, например Compuserve GIF;

далее на соответствующей позиции внутри документа делается графическая вставка.

Такой способ имеет ряд существенных недостатков: во-первых, вследствие использования большого количества различных программ, повышается время преобразования; во-вторых, поскольку графическая вставка имеет фиксированное разрешение, пользователь сталкивается с проблемой согласования шрифта с размером графической вставки; в-третьих, при печати документа на лазерном принтере шрифт вставки будет “зазубренным", несмотря на высокое разрешение печатающего устройства; в-четвертых, в связи с большим объемом документа получение его по компьютерной сети может занимать длительное время.

4. Inetrnet-технология подготовки математических текстов

Современные компьютерные технологии двигаются в сторону “усиления"серверных машин, стараясь разгрузить рабочие станции. Появились такие понятия, как “сетевой компьютер"и “тонкий клиент". Причиной этого является стремление достичь следующих целей: во-первых, снизить требования к аппаратной части рабочих станций; во-вторых, освободить пользователя от постоянных забот об обновлении программного обеспечения и поиска различных программ для поддержки разного рода сервиса. “Тонкие клиенты"имеют в своем составе только WWW-браузер, что вынуждает все больше ориентироваться на технологию работы с WWW-серверами. Основной принцип, провозглашенный Intranet-технологиями, — ориентация на “тонкого клиента".

С этой точки зрения, желательно иметь конвертор для создания электронных документов, который бы работал на клиентской машине, используя программное обеспечение сервера и создавал бы гипертекстовые документы математического содержания, не слишком перегруженные графическими файлами. Поэтому технология загрузки шрифтов, разработанная в Геометрическом центре университета Миннесоты, более других подходит для создания электронных публикаций математического содержания в среде Internet. Клиентское обеспечение, ориентированное на WWW, все больше использует Java-подобную технологию, что дает возможность легкого перехода на “сетевые компьютеры"и “тонких клиентов", а также позволяет сделать программное обеспечение независимым от платформы.

Исходя из этих соображений авторами статьи был разработан конвертор с TeXa в HTML, с представлением математических формул на основе технологии загрузки дополнительных шрифтов на машину клиента. Суть этой технологии заключается в том, что

перед просмотром математического текста на компьютер клиента загружаются математические шрифты (при повторном просмотре шрифты берутся уже из кэша и по сети не пересылаются), а отображение математических формул осуществляется Java-программой согласно их описанию (являющемуся расширением языка HTML и по структуре близкому к TEXовской нотации). Этот подход имеет следующие преимущества, по сравнению с традиционными:

отсутствие графических вставок делает текст публикации “читабельным"независимо от просмотрщика;

описание математических формул по своей нотации близко к описанию математических формул, принятому в TEXe, что существенно упрощает работу конвертора и наборщика;

если не считать объем шрифтов, то объем текстового файла такой же, как и при наборе в TEXe.

Преобразование документа из формата TEX в HTML производится программой, написанной на языке Java, и может выполняться как на машине клиента, так и на сервере. Особенностью программы является то, что математические выражения преобразуются к специальному виду, который интерпретируется при показе Java-приложением (applet), расширившим возможности HTML способностью поддерживать язык описания математических формул и математических акцентов, близких к формату TEX. В генерируемый файл автоматически включается вызов аплета для отображения математических выражений. Шрифты Java-аплетов могут загружаться с сервера (кстати, при повторном обращении они будут браться из кэша, что снижает нагрузку на сеть) или с жесткого диска клиента. Это позволяет при запуске приложений из памяти клиентской машины значительно повысить производительность.

Для работы клиента с этим пакетом достаточно компьютера с 386-м процессором и 8 Мб оперативной памяти под MS Windows-95 и 12 Мб свободного дискового пространства на системном диске для кэша. Эти требования являются минимальными для удовлетворительного функционирования MS Windows-95.

В разработанном конверторе вопросы, связанные с разбивкой TEXовского текста на отдельные WWW-страницы, установлением гиперсвязей между этими страницами и организацией навигации, остаются за клиентом. Мы сознательно отказались от автоматизации этого процесса в связи с тем, что электронная публикация по своей структуре и дизайну существенно отличается от печатной публикации и механический перенос принятых в печатных изданиях правил навигации (содержание, предметный указатель и др.) не всегда является удобным для публикации электронной.

Поступила в редакцию 24 апреля 1997 г.