2001 - Том 4 - Выпуск 2

Новые технологии компрессии для представления печатных документов в Интернет

Богдан Смолка (Bogdan SMOLKA)
Конрад Войцеховский (Konrad WOJCIECHOWSKI)
Технический Университет Силезии, отдел компьютерных технологий
(Silesian Technical University, Department of Computer Science)
Адрес: Akademicka 16, 44-101 Gliwice, POLAND
E-mail: bsmolka@ia.polsl.gliwice.pl

Большая часть культурного наследия сейчас доступна только на бумаге. Архивы, библиотеки, музеи – это места хранения большей части научного и культурного наследия. К сожалению, прямой доступ к этому огромному массиву материалов, бесценных для профессионального и личного использования, чрезвычайно затруднен и дорог.

Вот почему для того, чтобы обеспечить сохранность информации и простой доступ к ней студентам и ученым, наиболее интересные документы были пересняты. Микрофильмы дают неплохое качество, длительное хранение, незначительную зависимость от технических средств и возможность производить дополнительное копирование без значительных потерь информации. Однако, для того, чтобы обеспечить удаленный доступ к информации, микрофильмы должны быть оцифрованы, что означает дополнительную потерю качества.

В период быстрого развития коммуникационных и информационных технологий с все возрастающими возможностями передачи информации микрофильмы остаются в старой аналоговой эре. Новая эра принадлежит Интернету и быстрый рост технических преимуществ технологий оцифровки. Взрывообразное развитие Интернета как универсальной платформы обмена информацией предоставляет возможность легкого доступа к сокровищам культуры.

Однако, электронный доступ к печатным или письменным документам – сложная задача. Бумажные документы содержат текст и иллюстрации, и очень часто необходимо оцифровывать и переводить в формат изображения весь документ. Иногда возможно использовать технологии оптического распознавания текста и вычленять текст из рисунков и фотографий, но это представляется слишком сложным, когда имеешь дело со старыми документами, а во многих случаях и вообще невозможным. Старый документ не является суммой информации, содержащейся в тесте и в иллюстрациях. Цвет, текстура бумаги, стиль почерка или техника печати очень часто более важны, чем информационное содержание документа, которое в большинстве случаев уже известно.

Философия, лежащая в основе этого подхода, состоит в том, что старые документы необходимо представлять интегрировано, показывая текстовую

информацию в визуальном контексте. Для этого документы оцифровываются и представляются как изображения. Так называемые виртуальные библиотеки дают возможность просматривать документы, копировать их с помощью принтера и, что также важно, документы можно собирать и хранить в личном архиве для дальнейшего использования.

Главная проблема при представлении печатных документов в Интернет состоит в компромиссе между качеством воспроизведения документа и временем, необходимым для передачи и загрузки огромного объема информации, содержащейся в файле изображения.

Сейчас быстрый рост количества пользователей Интернет осложняет задачу, так как возможности передачи почти исчерпаны. Возможно, с внедрением быстрого Интернета проблема будет смягчена, но сейчас и в ближайшем будущем единственно возможным решением может быть компрессия изображения с потерями качества, позволяющая пользователю получить доступ к документу в разумное время с приемлемым качеством изображения. В результате оцифровки документа обычного размера со средним разрешением и глубиной цветопередачи чаще всего получается огромный файл размером порядка 20 – 50 Mb. Качество такого изображения высокое, но время передачи и обработки файла, а также технические требования к компьютеру непомерны. Вот почему повсеместно используется компрессия с потерей качества, так как она позволяет уменьшить размер файла при небольшой потере качества.

Наиболее популярные стандарты для передачи изображений форматы GIF и JPEG.

Формат GIF обычно используется для сжатия изображений, содержащих небольшое количество различных цветов. Так как этот формат использует схему кодирования без потерь качества, он не очень эффективен и не подходит для распространения реального живого цвета или изображений в серой шкале цвета. Гораздо лучшие результаты можно получить с помощью формата JPEG, разработанного Joint Pictures Expert Group, который осуществляет разделение хроматической информации, квантование трансформационных коэффициентов DCT и Huffman кодирование данных. Хотя коэффициент компрессии порядка 40:1 легко достигается без большой потери качества, формат JPEG не подходит для компрессии документов. Так как документ содержит много высокочастотных объектов, таких как буквы и рисунки, элиминирование высокочастотных компонентов при трансформации ведет к существенным потерям качества при воспроизведении документа. При повышении коэффициента компрессии текст быстро искажается и становится неразборчивым. Чтобы сделать документ читаемым необходимо делать файл JPEG большого размера, и это основное препятствие при создании эффективной Интернет-библиотеки.

Обычно принимается решение о переводе документа в двуцветный вид, а затем применяется компрессия стандартом CCITT, разработанным Fax Group 3 или Fax Group 4. Этот подход обеспечивает разборчивость текста при большом коэффициенте сжатия ценой полной потери информации о цвете.

JPEG, GIF и факсовые форматы, используемые для распространения документов,

сейчас заменяются новым форматами, основанными на волновом принципе, направленном на прямую компрессию документов, отсканированных с высоким качеством. Эти новые форматы обеспечивают быструю передачу оцифрованных документов при приемлемом уровне качества.

Среди новых волновых форматов три представляют наибольший интерес для представления оцифрованных документов в Интернете, это форматы DjVu, LuraDocument, и MrSID.

О Формат DjVu* разработан AT&T Bell Labs с целью сконструировать формат, который обеспечит распространение оцифрованных документов высокого разрешения по сетям. DjVu представляет изображение, используя 3 слоя:

· маска, которая является двуцветным представлением текста и рисунков. Этот слой, сжимаемый без потерь, показывает, какой пиксель относится к переднему плану (текст, рисунки, подписи), а какой - к фону (текстура бумаги, фотография).

· второй уровень представляет цвета фона, используя трансформацию, основанную на волновом кодировании.

· третий уровень содержит информацию о переднем плане изображения, закодированную с помощью того же самого волнового алгоритма. Используя DjVu, возможно точно воспроизвести документ, отсканированный с разрешением 300 dpi, уменьшив размер файла с 25 MB до 100 - 200 kB. Качество изображений в формате DjVu приемлемое, и этот формат подходит для компрессии страниц книг, газет, цветных фотографий, каталогов и т.д. Размер сжатого изображения документа настолько небольшой, что он подходит для распространения документов на CD-ROM (на одном CD-ROM можно разместить примерно 5000 газетных страниц) и через Internet (Internet-страница содержит в среднем около 50 KB информации). Еще одна важная характеристика формата DjVu – наличие удобных и бесплатных Internet- броузеров, которые позволяют быстро просматривать изображения, изменять их размеры, переводить в черно-белый формат и предоставляют много других полезных возможностей.

О Так как электронные изображения и сканированные документы требуют огромную память, возможности для хранения и большую скорость передачи, компания LuraTech ** разработала собственный стандарт для компрессии изображений LuraDocument (LDF), основанный на новейших достижениях волновой теории.

LDF, так же, как и DjVu, разделяет документ на текст и фон. Такой подход позволяет достичь высокого коэффициента компрессии. Используя волновой алгоритм сжатия текста и фона, формат LDF обеспечивает сжатие приблизительно 200:1, сохраняя приемлемое качество изображения. Этот формат предназначен специально для обработки изображений документов и дает возможность распространять их по глобальным и локальным сетям и использовать в коммерческих и некоммерческих приложениях. Формат LDF дает значительную экономию дисковых и сетевых ресурсов и может быть использован на многих компьютерных платформах. Также разработаны эффективные средства просмотра

изображений для популярных Internet-броузеров.

О LizardTech's MrSID * для фотографий – формат для кодировки больших изображений высокого разрешения, уменьшающий первоначальные размеры файла при сохранении высокого качества изображения. Изображения легко просматривать, уменьшать или увеличивать размеры, печатать без больших потерь качества. MrSID разработан специально для сжатия огромных файлов сканированных документов, старых книг, газет, особенно больших географических карт. Этот независимый от аппаратуры формат изображений дает оптимальное разрешение для экрана и для Internet.

Этот формат дает возможность быстро переслать через Internet изображение документа, готовое для печати, трансформация файла требует нескольких секунд. То же самое верно и для процесса кодирования. MrSID требует всего несколько секунд для открытия больших файлов. MrSID дает хороший коэффициент сжатия без заметной потери качества.

Одно из преимуществ формата MrSID – простота использования. Можно сжимать изображение из Photoshop® или использовать MrSID Workgroup Encoder. Файлы MrSID поддерживаются обычно используемыми графическими редакторами и стандартными Web-броузерами, что дает пользователям возможность гибкого использования MrSID при оцифровке своих коллекций.

Как показано в этом коротком докладе, новые технологии компрессии, основанные на волновом алгоритме, дают значительное увеличение качества по сравнению со стандартными форматами и позволяют представлять сканированные документы в Internet. Высокое качество сжатых изображений и в среднем небольшие размеры позволяют использовать их для представления виртуальных библиотек в глобальных сетях.

Наш опыт использования новых технологий показывает, что их необходимо использовать повсеместно для того, чтобы предоставлять легкий и быстрый доступ к материалу высокого качества, доступного только в бумажной форме. Внедрение новых технологий безусловно продвинет Internet на шаг вперед по пути превращения в универсальное, наиболее мощное средство обмена информацией.

Возможности новых форматов можно посмотреть на экспериментальном WWW-сайте http://plum.ia.polsl.gliwice.pl/vb, подготовленном вместе с Silesian Library, Katowice, Poland. На этом сайте мы собрали много примеров разных документов, таких.

как инкунабулы, старые письма, газеты, сборники песен, фотографии, карты. Каждый документ представлен в формате DjVu и в формате LuraDocument, для того, чтобы их можно было сравнивать. Наша виртуальная библиотека предоставляет информацию и о других проектах, целью которых является доступ к культурному наследию через Internet.

(перевод Н. Браккер)

Богдан Смолка работает в отделе автоматического управления Университета. Родился в 1962 году, закончил университет в 1986 году по специальности Физика, доктор технических наук (1998).
Занимается обработкой изображений (контрастность, уменьшение шума, реставрация), компрессией изображений (волновые технологии компрессии) и восприятием качества изображения. В настоящее время его научные интересы включают представление старых документов в Интернет с помощью новых эффективных технологий компрессии.

Технологический Университет Силезии был основан по декрету Президента Национального народного совета 24 мая 1945 года после длительных попыток открыть технический университет в районе Верхней Силезии, предпринимавшихся с 1920 г.

В 1945 году основной профессорско-преподавательского состава были бывшие профессора Технического Университета г. Львова. Многочисленные профессора Университета в последующие периоды тоже были выходцами из Технического Университета г. Львова, поэтому учебные планы и программы составлялись на основе опыта львовского Университета. В последующие годы тесные связи двух университетов укреплялись. Прекрасный преподавательский состав был основным преимуществом вновь созданного Технологического Университета Силезии, что отличало его от других польских университетов.

В настоящее время Технологический Университет Силезии – это большой и современный университет, в котором обучается около 25000 студентов. В нем читаются курсы по 21 инженерной специальности, в том числе дневные и вечерние курсы для получения степени магистра и бакалавра, а также дополнительные курсы на степень магистра. Предлагаются дополнительные аспирантские курсы и курсы для получения степени доктора технических наук по наиболее привлекательным дисциплинам и становятся все более и более популярными. Некоторые исследования и курсы преподаются на английском и французском языках.

Изображения университетских зданий и лабораторий можно найти на http://www.polsl.gliwice.pl/alma.mater/pictures.html