в) з назвами днів тижня і датами.

It was extraordinary that after thirty years of marriage his wife could not be ready in time on Sunday morning.

У даному прикладі, ми бачимо, що після тридцяти років шлюбу дружина священика не могла бути готова вчасно саме в недільний ранок, тобто прийменник on в даному контексті використовується при вказівці дня тижня. Сполучниками називаються службові слова, що служать для з'єднання членів пропозиції в простому реченні і простих речень у складному реченні. Сполучники можуть бути використані при вказівці часу в реченні.

Uncle William used to tell Philip that when he was a. curate his wife had known twelve songs by heart, which she could sing at a moment's notice whenever she was asked.

У даному прикладі за допомогою союзу whenever ми бачимо, що дружина священика знала дванадцять пісень напам'ять і могла заспівати їх всякий раз, коли б її не запитали.

But Philip could not bear to be angry with him long, and even when convinced that he was in the right, would apologise humbly. У даному контексті за допомогою сполучника when ми можемо бачити, що Філій не мі)' довго злитися, навіть тоді, коли був переконаний в своїй правоті.

Отже ми дійшли **висновку**, що на кожному рівні виділяються одиниці, які мають темпоральне значення. Так, на морфологічному рівні такими одиницями є певні граматичні форми, на лексичному - лексичні одиниці з гемпоральним значенням, на рівні речення -• синтаксичні конструкції . Часові відношення виражаються в семантиці мовних одиниць на граматичному й лексичному рівнях. Поряд з граматичними субкатегоріями темпоральності для закріплення в мові цих уявлень формуються особливі лексичні одиниці, що утворюють домінуючі лексико-семантичні поля, котрі вказують на специфіку усвідомлення світу, на особливості орієнтації того чи іншого народу в часі.

Надійшла до редколегії 17.06.2014.

УДК 004.912:004.056.52:517.5

ШУМЕЙКО О.О., д.т.н., професор
ТИМОШЕНКО Д.В., к.т.н.

Дніпродзержинський державний технічний університет

ВИКОРИСТАННЯ ЗАХИЩЕНОГО ФОРМАТУ ДЛЯ ОБМЕЖЕННЯ НЕСАНКЦІОНОВАНОГО РОЗПОВСЮДЖЕННЯ ДОКУМЕНТІВ

Вступ. Враховуючи сучасний розвиток інформаційних технологій, усе більша значимість надається інформації, що отримується через Інтернет або локальну мережу. Перш за все це публічні електронні бібліотеки, музеї, архіви, ресурси для дистанційної освіти. У даному контексті актуальним є впровадження нових інформаційних технологій автоматизованої обробки електронних документів та їх передачі через глобальну чи локальну мережу, які б дозволили ефективно використовувати існуючи інформаційні ресурси.

Обробка та передача електронних документів має свою специфіку. Особливо, якщо це стосується растрових електронних документів, тобто коли насправді документ не є звичайним текстом, а має вигляд зображення. У світі існують декілька спеціалізованих форматів, орієнтованих на зберігання растрових електронних документів: DJVU, JPEG2000/Part 6 та LuraDocument. В основу цих форматів покладено ідеї щодо розбиття

документа на кольорові прошарки, смислову інформацію та фон. За своєю суттю ці формати є локальними, через що їх використання у мережі зводиться до передачі файлу в цілому. Фактично передається копія документа і подальше її використання дуже складно проконтролювати. Отже, електронна бібліотека у такому вигляді за функціонуванням суттєво відрізняється від стандартної бібліотеки: замість представлення тільки інформації, вона назавжди надає носій всієї інформації, що у подальшому може бути використаним незалежно від самої бібліотеки. Такий підхід входить у протиріччя із законодавством щодо захисту прав інтелектуальної власності (Закон України "Про авторське право і суміжні права" від 23.12.1993 № 3792-XII).

Внаслідок вимог чинного законодавства до частини авторських документів взагалі не існує публічного доступу через проблему несанкціонованого та неконтрольованого розповсюдження. Разом з тим, при наданні певного ступеня складності такого розповсюдження до такого роду документів було б відкрито доступ для ознайомлення.

Значний розвиток цифрової обробки зображень, яка лежить в основі обробки растрових електронних документів, зробили вітчизняні та іноземні вчені, як П. Хафнер, Л. Ботоу, П. Ховард, Я. Ле Кун, М.П. Корнійчук, А.О. Лигун, О.О. Шумейко та інші.

Таким чином, розробка сучасних методів обробки, стиску та передачі електронних документів, які дозволяють організувати публічний доступ до інформації з обмеженням несанкціонованого розповсюдження документів, є актуальною науково-технічною задачею.

Постановка задачі. Пропонується сконцентрувати увагу на вирішенні двох проблем:

1. Організація формату електронного растрового документу та взаємодію з ним за участі клієнт-серверної архітектуру для контролю за доступом до інформації.
2. Покращення якості стиску електронного растрового документу за допомогою врахування специфіки документа.

Звичайно, ці задачі пов’язані одна з одною.

Почнемо розгляд пропозицій з другої задачі.

Результати роботи. Головна ідея полягає в розробці інформаційної моделі електронного растрового документу, яка обумовлює розподіл документа на 4 інформаційні прошарки: прошарок символів, прошарок «ділова графіка», прошарок зображень та фон. Більш детальне розбиття на прошарки дозволить більш ефективно застосувати відповідний апарат стиску.

Розглянемо запропоновані прошарки на прикладі рис.1.

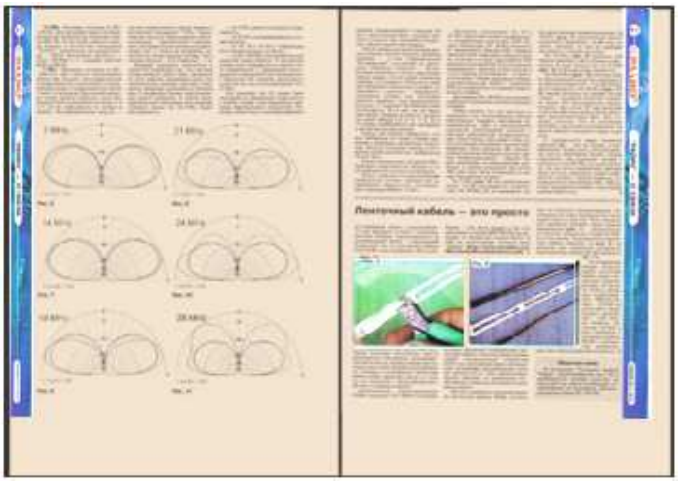

Рисунок 1 – Оригінальне зображення

1. Прошарок символів (рис.2). До такого прошарку належать всі текстові елементи, які зустрічаються в електронному растровому документі: символи, знаки пунктуації. Зазвичай, у подальшому будемо розуміти під символами всі елементи цього прошарку. Для цього прошарку характерні часті появи символів, які по суті є одними й тими ж. Ця властивість є базовою для стиску такого прошарку.

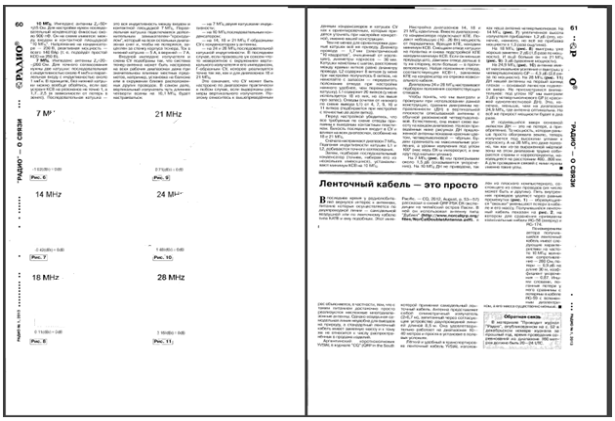

Рисунок 2 – Прошарок симовлів

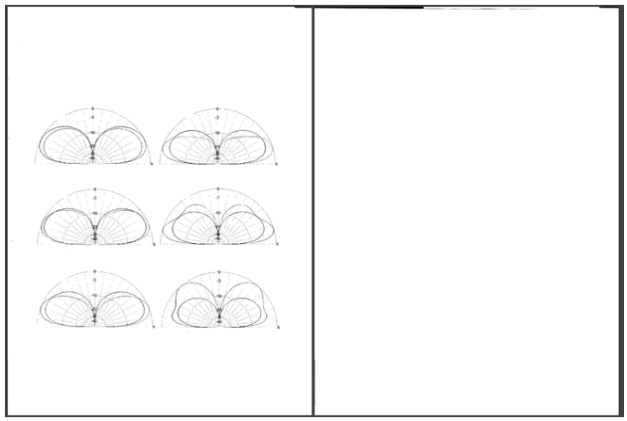

Рисунок 3 – Прошарок «ділова графіка»

2. Прошарок «ділова графіка» (рис.3). До цього прошарку відносяться лінії, діаграми, таблиці. Виходячи зі специфіки таких даних з точки зору їхнього стиску найбільш ефективно використовувати методи векторизації для того, щоб замість збереження бінарного виду прошарку, зберігати основні вузли елементів і відновлювати за допомогою відповідних апаратів наближення.

3. Прошарок зображень (рис.4). Цей прошарок складається з зображень, які зустрічаються у тексті. Це можуть бути, наприклад, комікси, ілюстрації. Для стиску елементів цього прошарку використовуються методи стиску зображень зі втратами або без втрат.

4. Прошарок фону (рис.5). В залежності від того, як виглядає фон – це просто білий колір або текстура, або зображення, наприклад, як в журналах, до нього можна використовувати різні методи стиску. У загальному випадку фон стискається як звичайне зображення з більш високим показником стиску.

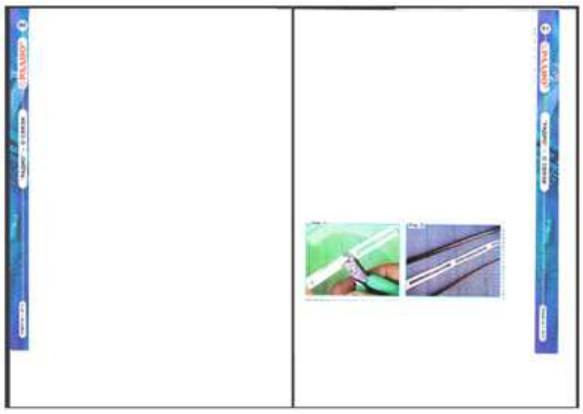

Рисунок 4 – Прошарок зображень

Рисунок 5 – Прошарок фону

Таким чином, подальші міркування присвячені вилученню з документів прошарків та методам їхньої обробки та стиску.

Проблеми, з якими ми зустрінемося при локалізації символів: невизначеність кольору символів та фону (може бути и чорний текст на білому фоні, як зазвичай, а може і навпаки – білий текст на чорному фоні), сам фон може бути текстурованим або й зовсім мати вигляд зображення, символи можуть належати різним мовам та різним шрифтам. Ці проблеми можна вирішувати за рахунок оригінального адаптивного алгоритму, що базуєтьтся на званому квантуванні Ллойда-Макса.

Не менш складна проблема, що пов'язана з обробкою прошарку символів, – кластеризація. Кластеризація необхідна для стиску цього прошарку. Кластеризацію пропонується робити за участі квантування Ллойда-Макса для первинного розбиття на кластери та методу головних компонент для уточнення елементів кластеру та побудови шаблону.

Для локалізації більшості ліній в прошарку «ділова графіка» пропонується використовувати фільтр Кенні, але він не є дуже вдалим для локалізації ліній, що слабо вирізняються на навколишньому фоні. Для локалізації таких ліній пропонується розробити новий алгоритм, який буде врахувати слабку відмінність точок, що належатимуть таким ліням, від фону. Для описую елементів «ділової графіки» пропонується використовувати аналітичне представлення прямої, кола, сплайнів. Це дасть можливість зберігати тільки принципові точки, тип апарату наближення, колір та товщину.

Зберігання прошарку зображень та фону можна проводити за участі, наприклад, формату JPPEG 2000.

Як вже було сказано вище необхідно розробити новий формат даних для організації контролю за розповсюдженням інформації. Спираючись на новий формат можна виділити такі основні компоненти програмного комплексу:

1. Кодер послідовності зображень у спеціальний формат ALD.
2. Декодер формату ALD в послідовність зображень.
3. ALD-сервер для передачі даних клієнтові.
4. Плагін до Internet Explorer (як самого популярного браузера) для взаємодії з ALD-сервером та інтерпретації даних у вигляді зображення.

Формат ALD відображає інформаційну модель електронного растрового документу, яка разом зі способами обробки та стиску була розглянута у попередніх розділах. А саме, окремо виділені байти для збереження прошарку символів, прошарку «ділова графіка», прошарку зображень та прошарку фону.

Разом з тим, базуючись на тому, що електронний растровий документ – це набір зображень, кожне з яких відповідає сторінці цього документу, специфічною властивістю формату ALD є те, що він надає можливість розкриття даних тільки на необхідній сторінці (сторінках), у зв'язку з чим контейнер має наступний вигляд (рис.6):

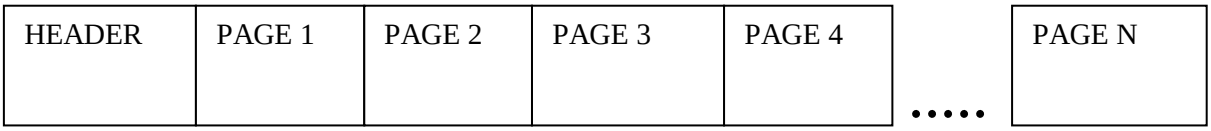

Рисунок 6 – Схема збереження сторінок

При наявності заголовку файлу (HEADER) кожна сторінка відкривається незалежно від інших. Наприклад, із вже існуючого файлу можна сформувати робочий документ такого роду (рис.7).

Така структура дала можливість використовувати даний формат в Інтернеті. Користувач сам визначає (можливо, виходячи зі змісту) ті сторінки, які йому потрібні, і відповідно тільки їх і завантажує. Варіантом, що забезпечує повну незалежність всіх сторінок одна від одної, був би варіант з повною незалежною обробкою (стиском) кожної сторінки. Але в такому ви-

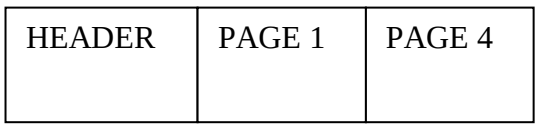

Рисунок 7 – Схема вибіркового відкриття сторінок

падку ступінь стиснення документа знизилася б, і користувач відчув би це більшою платою за кожну завантажену сторінку зі збільшенням їх кількості. Протилежний варіант – обробляти всі сторінки цілком. Дійсно, ступінь стиснення має більший високий рівень, але користувач, завантажуючи одну сторінку, автоматично завантажує всю інформацію, що стосується решти сторінок. Знову ж такий варіант відіб'ється на платні за цю послугу, і тим сильніше, чим менше число сторінок потрібно користувачеві. Необхідно було знайти компроміс між двома крайніми варіантами. На даний момент він являє собою наступний варіант формату ALD (рис.8):

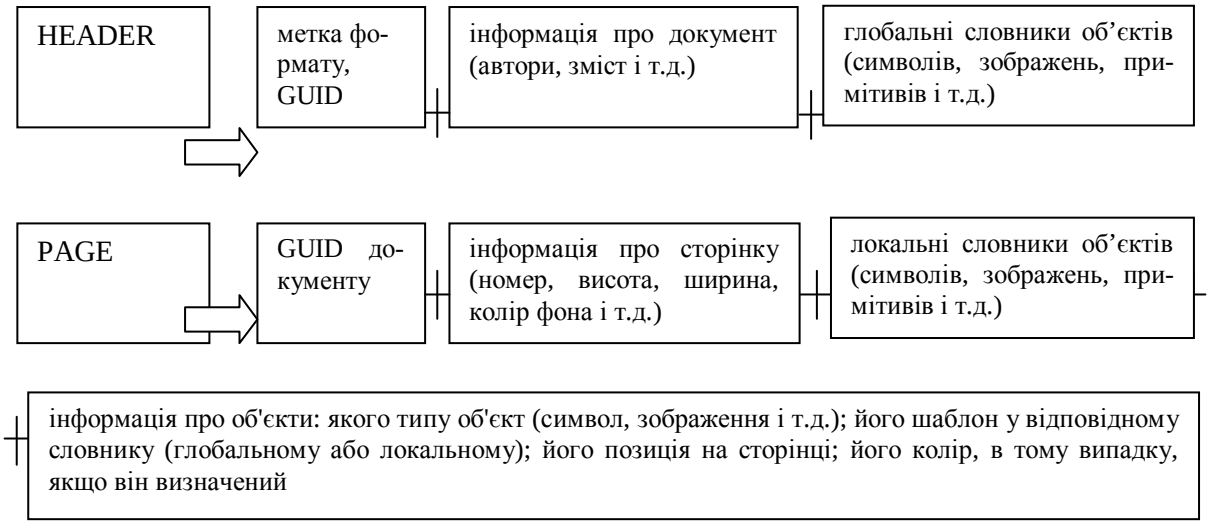

Рисунок 8 – Схема складових HEADER, PAGE

Висновки. Спираючись на схему 8, можна сформулювати такий алгоритм взаємодії сервера з клієнтом:

1. Користувач відкриває в браузері посилання, яке веде на документ у форматі ALD.
2. На сервер відправляється запит на базову інформацію (кількість сторінок, глобальний словник).
3. Браузер отримує базову інформацію, відображає доступні номери сторінок, користувач вибирає номер сторінки для перегляду.
4. На сервер відправляється запит на повну інформацію для конкретної сторінки.
5. Браузер отримує бінарні дані, за допомогою встановленого декодера, перетворює дані в зображення і відтворює у вікні.

ЛІТЕРАТУРА

1. Лигун А.О. ALLDocument – технологія нового покоління для збереження, передачі та відображення електронних документів / А.О.Лигун, О.О.Шумейко, Д.В.Тимошенко // Вісник Східноукраїнського національного університету імені Володимира Даля. − 2006. − №9 (103). – Частина 1. − С. 83-85.
2. DjVu: Analyzing and Compressing Scanned Documents for Internet Distribution /Patrick Haffner, Léon Bottou, Paul Howard, Yann Le Cun // Proceedings of the International Conference on Document Analysis and Recognition. – 1999. – P. 625-628.
3. ISO/IEC 15444-6:2003 Information technology JPEG 2000 image coding system - Part 6: Compound image file format. – 2003. – 71p.

Надійшла до редколегії 17.06.2014.