УДК 681.3

В. Левыкин, Е. Моспан
Харьковский национальный университет радиоэлектроники,
кафедра информационных управляющих систем

РАЗРАБОТКА УНИФИЦИРОВАННОЙ МОДЕЛИ ПРЕДСТАВЛЕНИЯ СТРУКТУР ЭЛЕКТРОННЫХ ДОКУМЕНТОВ

Описаны результаты исследования моделей представления структур электронных документов, форматом которых являются описательные языки разметки. На основании выявленных недостатков существующих моделей разработана унифицированная модель представления структур электронных документов, не зависящая от типа описательного языка разметки. Данная модель позволяет использовать произвольный язык разметки описательного типа в качестве формата документа-шаблона в технологиях формирования электронных документов.

In this article investigation results are outlined about models for presenting structures of electronic documents, which format belongs to descriptive markup languages. On the basis of obtained disadvantages in existing models unified model for presenting structures of electronic documents is developed, which is independent from the type of descriptive markup language. This model allows to use optional descriptive markup language as format for document-pattern in the electronic document generation technologies.

Введение

Применение электронных документов для представления результатов реализации функциональных задач информационных систем является распространенной практикой на сегодняшний день. Для этого применяют технологии формирования электронных документов [1]. Для формирования выходного документа большинство технологий использует документы-шаблоны, форматом которых являются описательные языки разметки. Такие документы-шаблоны обеспечивают представление общей структуры выходного электронного документа, создание которых происходит в соответствующих текстовых процессорах. Модификация структуры документа-шаблона осуществляется на основании данных функциональной задачи, благодаря чему обеспечивается формирование электронного документа, отражающего эти данные. В работе [2] рассмотрена технология формирования электронных документов, которая предполагает использование произвольного описательного языка разметки в качестве формата документа-шаблона. Следовательно, задача исследования моделей представления структуры электронных документов, форматом которых являются описательные языки разметки, является актуальной.

Для представления структуры электронного документа в языках разметки описательного типа используются иерархические объектные модели. Такие модели позволяют исследовать элементы структуры электронного документа, как в отдельности, так и в целом [3]. При этом модели представления структур документов вытекают непосредственно из стандартов или спецификаций языков разметок. Доступ к элементам таких моделей обеспечивается посредством соответствующего текстового процессора или программного модуля. Так, например, при помощи технологии Component Object Model (COM) в операционной системе Windows существует возможность организовать работу программным способом с электронными документами, которые представлены форматами, поддерживаемыми Microsoft Word (RTF, DOC). Структуру документов в формате языков разметки HTML и OpenDocument можно представить при помощи DOM-модели,

поскольку они являются реализациями стандарта SGML и XML-совместимыми документами[4-5]. Благодаря таким моделям представления структур электронных документов и рассмотренным программным модулям можно обеспечить доступ к элементам структуры документов, динамически изменяя её содержимое. Однако указанные выше модели представления структур электронных документов неоднородны и предназначены для языков разметки определенного типа.

Обобщенную модель представления структуры электронных документов, форматом которых являются описательные языки разметки, представим ориентированным графом следующего вида:

$$E = (\{V_i\}, \{A_j\}), i = \overline{1,n}; j = \overline{1,m}, \quad (1)$$

где E – ориентированный граф, отражающий структуру электронного документа в формате описательного языка разметки; V_i – вершина ориентированного графа E; A_j – ребро ориентированного графа E; n – количество вершин ориентированного графа E; m- количество ребер ориентированного графа E.

Следует отметить, что вершинам V_i графа E соответствуют элементы структуры электронного документа в контексте некоторого описательного языка разметки, а ребро A_j определяет отношения между ними. Граф E представим следующим образом:

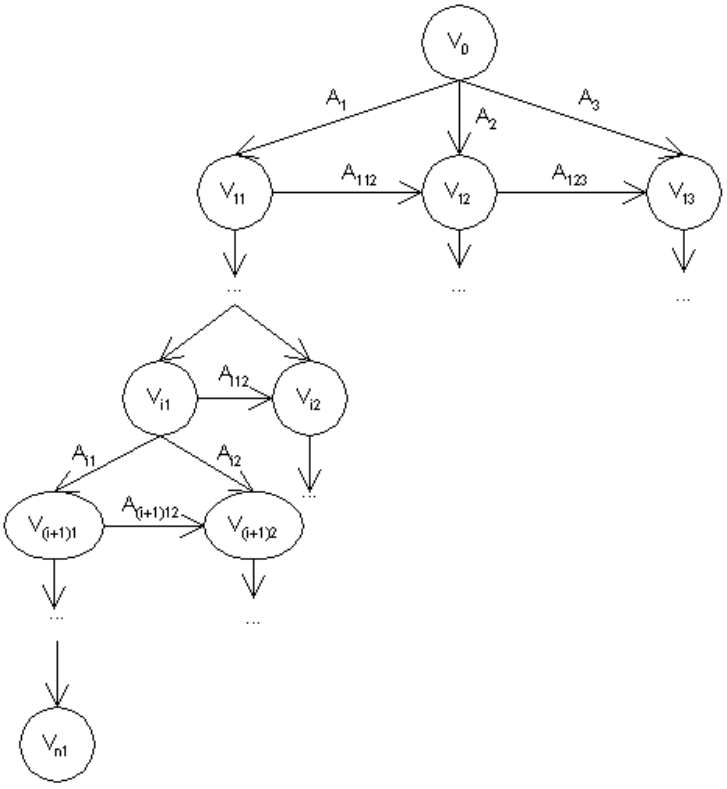

Рис. 1. Ориентированный граф E, который представляет структуру электронного документа

Ориентированный граф E имеет только один корень, а уровень его вершин определяется кратчайшим путем до корня. Вершине нулевого уровня (корню), как правило, соответствует объект структуры электронного документа, который представляет его в целом. Вершины первого уровня соответствуют объектам, которые представляют собой метаинформацию о страницах, составляющих электронный документ. Вершины, располагающиеся в графе E со второго по последний (n) уровень, соответствуют различным объектам, которые входят в состав структуры электронного документа. Их тип, последовательность, а также состав их свойств определяется спецификацией того описательного языка разметки, в формате которого представлен рассматриваемый электронный документ.

Постановка задачи

На основании проведенного выше анализа можно утверждать, что существующие модели представления структур электронных документов, которые имеют формат описательных языков разметки, являются узкоспециализированными. В связи с этим их невозможно использовать в технологии формирования электронных документов [2], которая предполагает использование в качестве формата документа-шаблона произвольные описательные языки разметки. Следовательно, разработка унифицированной модели представления структуры электронного документа, форматом которого является произвольный описательный язык разметки, является актуальной задачей. Такая модель должна удовлетворять следующим требованиям:

- модель должна представлять структуру электронного документа, который имеет формат произвольного языка разметки описательного типа;
- модель должна отражать специфические свойства любого объекта из спецификации произвольного языка разметки описательного типа;
- модель должна обеспечивать возможность разработки программного модуля для обеспечения гибкого изменения структуры электронного документа на основании актуального состояния данных некоторой функциональной задачи ИС.

Разработка унифицированной модели представления структур электронных документов

Структура электронного документа, форматом которого является описательный язык разметки, в общем виде может быть представлена следующим образом:

$$M = < \overline{E}, \overline{F} >, \quad (2)$$

где M – модель представления структуры электронного документа; $\overline{E}$ – множество объектов, определяющих эту структуру; $\overline{F}$ –множество отношений между объектами множества $\overline{E}$.

Для описания структуры электронного документа, форматом которого является описательный язык разметки, в контексте унифицированной модели необходимо определить множества $\overline{E}$ и $\overline{F}$. Исходя из вышеуказанных предпосылок к разработке унифицированной, модели можно сделать вывод о том, что состав множеств объектов и отношений модели M', определяется структурой электронного документа. Такую структуру документа представляет ориентированный граф E, описанный выше. Каждой вершине графа E соответствует некоторый объект структуры электронного документа, который определен в спецификации описательного языка разметки. Следовательно, множество объектов модели M' определяется на основании всего множества элементов, составляющих структуру электронного документа, которая представлена графом E. На основании проведенного анализа существующих моделей можно выделить два основных типа объектов, из которых будет состоять структура электронного документа, представленного унифицированной моделью. Первым типом являются логические объекты (E_l), которые соответствуют объектам-контейнерам. В рамках существующих описательных языков разметки логическим объектам соответствуют такие объекты как параграф (абзац), таблица, ячейка таблицы, сноска и другие. Вторым типом объектов являются физические объекты (E_p). Такие объекты соответствуют узлам каждой ветви, которая присутствует в графе. В рамках существующих описательных языков разметки физическим объектам соответствуют такие объекты, как текст, рисунки и другие.

Для того чтобы обеспечить возможность внесения изменений в структуру документа, представленного унифицированной моделью, необходимо определить тот набор объектов, который может подвергаться подобным действиям. Для этого необходимо ввести еще один тип объектов, которые могут входить в структуру документа в контексте унифицированной модели. Это, так называемые, квази-объекты (E_q). Основной особенностью квази-объектов является то, что они не оказывают влияние на визуализацию электронного документа в текстовом процессоре, а служат лишь метками некоторого пространства объектов, которые могут быть заменены, удалены или повторены в зависимости от актуального состояния данных функциональной задачи, в рамках

которой создается электронный документ. В общем виде квази-объекты можно представить следующим образом:

$$E_q =< T_E, f(E_d, T_E) >, \quad (3)$$

где T_E – множество объектов, которые относятся к квази-объекту E_q; $f(E_d, T_E)$ – функция, благодаря которой происходит изменение структуры электронного документа. E_d – объект, который отображает структуру электронного документа в целом в рамках унифицированной модели;

Физические объекты E_p определяют те данные, которые содержатся в электронном документе. В общем виде их можно представить следующим образом:

$$E_p =< TYPE, VALUE >, \quad (4)$$

где TYPE – тип физического объекта (текст, картинка и т.д.); VALUE – формализованный вид представления информации, которую содержит физический элемент E_p.

Если физические объекты определяют содержание электронного документа, то логические объекты E_l предопределяют его структуру и форматирование элементов документа, которое осуществляется посредством его визуализации в соответствующем тестовом процессоре. Логические объекты E_l соответствуют объектам языка разметки описательного типа, в формате которого представлен исходный электронный документ. Эти объекты обладают уникальным конечным набором свойств, которые определяют правила визуализации их в текстовом процессоре. Так, например, такой элемент, как параграф (абзац), обладает следующими свойствами: высота строки, отступы первой и последующих строк, свойства, которые относятся к его положению на странице. Для унификации описания свойств логических объектов введем специальный контейнер свойств, который далее будем называть стилем. Исходя из вышесказанного, представим логический объект E_l следующим образом:

$$E_l =< T_C, S >, \quad (5)$$

где T_C – объекты структуры электронного документа, для которых логический объект E_l является контейнером; S – стиль (контейнер свойств) логического объекта E_l.

Стиль S представим следующим образом:

$$S =< p_1, p_2, ..., p_i, ..., p_{n-1}, p_n >, \quad (6)$$

где p_i – свойство стиля S; n – количество свойств в стиле S.

Свойство p_i опишем кортежем вида:

$$p_i =< K, W >, \quad (7)$$

где K – ключ (название) свойства p_i; W – значение свойства p_i.

В контексте унифицированной модели логические объекты имеют некоторые разновидности, которые позволяют представить специфические объекты-контейнеры. Такие разновидности соответствуют вершинам определенного уровня в контексте графа E. Так корню графа соответствует объект E_d, который является разновидностью логического элемента и представляет собой документ в целом. Стиль объекта E_d содержит в себе метаинформацию о документе (например, сведение об авторах, назначении документа и прочее). Объекты, соответствующие вершинам 1-го уровня, представляют собой секции электронного документа. Под секцией следует понимать объект, который хранит в себе метаинформацию о размерах страницы, отступах до текста, колонтитулах и т.д. Объект, соответствующий секции E_s в унифицированной модели представим следующим образом:

$$E_s =< H, F, SB, S >, \quad (8)$$

где H – логический объект, соответствующий верхнему колонтитулу; F – логический объект, соответствующий нижнему колонтитулу; SB – основной логический объект секции, который

является контейнером для всех объектов структуры электронного документа, образующие тело секции; S – стиль логического объекта E_s, который содержит метаинформацию о свойствах секции, определяемую спецификацией описательного языка разметки.

Следующей разновидностью логических объектов в контексте унифицированной модели представления структур электронного документа являются контейнеры физических объектов. Обозначим такие объекты как E_{pc}. Их характерной особенностью является то, что они являются контейнером только для одного физического объекта. В рамках своего стиля объекты E_{pc} содержат свойства о дочернем физическом объекте. Так, например, для физического объекта типа «ТЕКСТ» стиль объекта E_{pc} обладает свойствами о шрифте (типе, размере и гарнитуре), на основании которого он должен быть визуализирован в текстовом процессоре. В общем виде объект E_{pc} можно представить следующим образом:

$$E_{pc} =< E_p, S, C >, \tag{9}$$

где E_p – дочерний физический объект; S – стиль объекта E_{pc}; C – тип связи со следующим контейнером физических элементов, который определяет взаиморасположение двух рядом расположенных объектов E_{pc}.

Возможные типы связи C определяются спецификацией описательного языка разметки. Например, в языке разметки RTF существуют такие типы связи: пробел, неразрывный пробел, знак табуляции, мягкий перенос и другие.

На основании выше сказанного представим унифицированную модель M' представления структуры электронного документа, форматом которого является произвольный описательный язык разметки, моделью следующего вида:

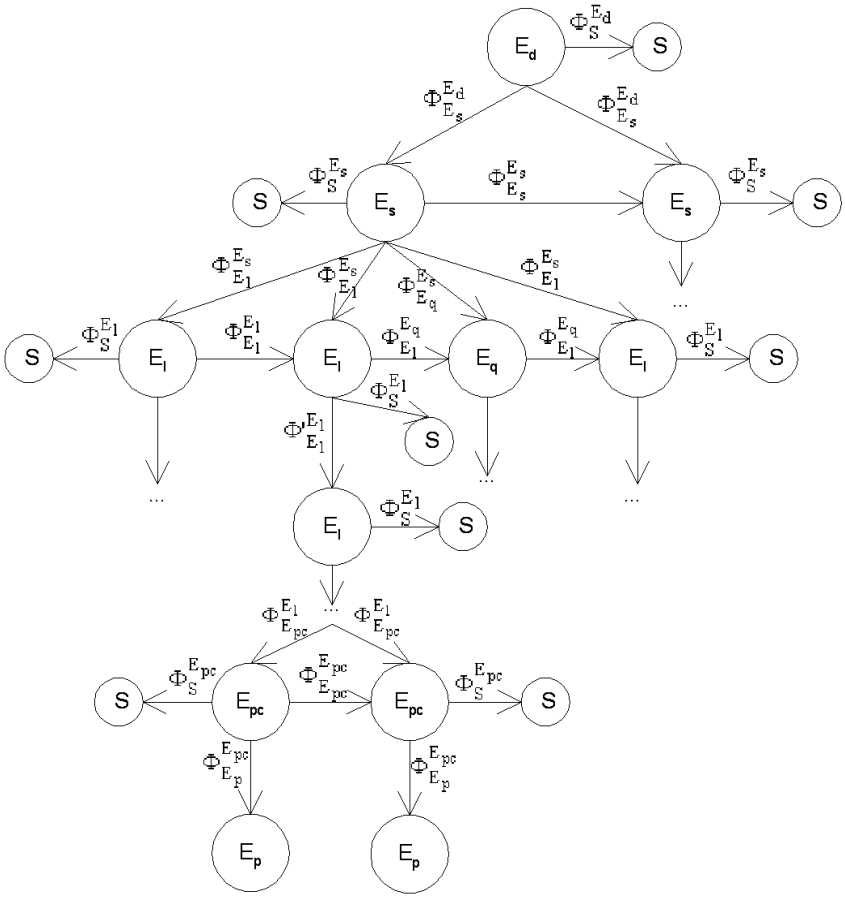

Рис. 2. Структура электронного документа в контексте унифицированной модели M'

Запишем унифицированную модель M' в общем виде:

$$M'=< E_d, \overline{E_s}, \overline{E_l}, \overline{E_q}, \overline{E_{pc}}, \overline{E_p}, \overline{S}, \Phi_S^{E_d}, \overline{\Phi_{E_s}^{E_d}}, \overline{\Phi_{E_s}^{E_s}}, \overline{\Phi_S^{E_s}}, \overline{\Phi_{E_l}^{E_s}}, \overline{\Phi_{E_l}^{E_l}}, \overline{\Phi'^{E_l}_{E_l}}, \overline{\Phi_S^{E_l}}, \overline{\Phi_{E_q}^{E_s}}, \overline{\Phi_{E_l}^{E_q}}, \overline{\Phi_{E_{pc}}^{E_l}}, \overline{\Phi_{E_{pc}}^{E_{pc}}}, \overline{\Phi_S^{E_{pc}}}, \overline{\Phi_{E_p}^{E_{pc}}} > \quad (10)$$

где M' – унифицированная модель представления электронного документа; E_d – объект, представляющий документ в целом; $\overline{E_s}$ – множество объектов E_s, представляющих секции документа; $\overline{E_l}$ – множество логических объектов E_l; $\overline{E_q}$ – множество квази-объектов E_q; $\overline{E_{pc}}$ – множество контейнеров физических объектов E_{pc}; $\overline{E_p}$ – множество физических объектов E_p; $\overline{S}$ – множество стилей S; $\Phi_S^{E_d}$ – отношение между объектом E_d и стилем S, определяющее связь между ними; $\overline{\Phi_{E_s}^{E_d}}$ – множество отношений между объектами E_d и E_s, которое показывает, из каких секций состоит электронный документ; $\overline{\Phi_{E_s}^{E_s}}$ – множество отношений между двумя объектами E_s, определяющее порядок следования секций в документе; $\overline{\Phi_S^{E_s}}$ – множество отношений между объектами E_s и стилями S, определяющее связь между ними; $\overline{\Phi_{E_l}^{E_s}}$ – множество отношений между объектами E_s и E_l, которое показывает, из каких логических объектов E_l состоит секция E_s; $\overline{\Phi_{E_l}^{E_l}}$ – множество отношений между двумя объектами E_l, определяющее порядок их следования в рамках общего контейнера; $\overline{\Phi'^{E_l}_{E_l}}$ – множество отношений между двумя объектами E_l, которое определяет, какие логические объекты E_l являются дочерними элементами другого логического объекта; $\overline{\Phi_S^{E_l}}$ – множество отношений между объектами E_l и стилями S, определяющее связь между ними; $\overline{\Phi_{E_q}^{E_s}}$ – множество отношений между объектами E_s и E_q, которое показывает, какие квази-объектов E_l входят в секцию E_s; $\overline{\Phi_{E_l}^{E_q}}$ – множество отношений между объектами E_l и E_q, определяющие порядок их следования в рамках общего контейнера; $\overline{\Phi_{E_{pc}}^{E_l}}$ – множество отношений между объектами E_l и E_{pc}, которое показывает, из каких контейнеров физических объектов E_{pc} состоит логический объект E_l; $\overline{\Phi_{E_{pc}}^{E_{pc}}}$ – множество отношений между двумя объектами E_{pc}, определяющее порядок их следования в рамках общего контейнера; $\overline{\Phi_S^{E_{pc}}}$ – множество отношений между объектами E_{pc} и стилями S, определяющее связь между ними; $\overline{\Phi_{E_p}^{E_{pc}}}$ – множество отношений между объектами E_{pc} и E_p, которое показывает, из каких физических объектов E_p состоит объект E_{pc};

Выводы

Разработанная унифицированная модель позволяет представлять структуры электронных документов, форматом которых является произвольный описательный язык разметки. Эта модель описывает иерархическую объектную структуру электронного документа и отражает свойства её элементов. Квази-объекты, которые входят в состав разработанной модели, позволяют модифицировать структуру электронного документа. Такие особенности унифицированной модели

определяют возможность ее применения в технологиях формирования электронных документов для представления структур документов-шаблонов, форматом которых является произвольный описательный язык разметки.

1. Разработка модели модифицированной технологии формирования электронных документов на основании шаблонов в WEB-ориентированных информационных системах / С.Ф. Чалый, Д.Л. Кравченко, Е.А. Моспан // Сборник научных трудов Харьковского университета воздушных сил. – 2008. – Вип. 3 (18). – С. 135–138. 2. Левыкин В. М., Моспан Е. А. Разработка модели формирования электронных документов в WEB-ориентированных информационных системах // АСУ и приборы автоматики. – 2008. – Вып. 144. – С. 54–58. 3. Markup Languages: Theory and Practice: Journal. – Cambridge: MIT Press, 1999. – 120 p. 4. Молли Э Хольцшлаг Использование HTML и XHTML: Using HTML and XHTML : Спец. изд.: Пер. с англ. – Издательский дом Вильямс, 2004, 728 с. 5. Charles F. Goldfarb, Yuri Rubinsky The SGML Handbook. – Oxford University Press, 1990, 663 p.

УДК 519.15:621

О. Різник, Б. Балич, І. Вербенко
Національний університет “Львівська політехніка”,
кафедра автоматизованих систем управління

ВИКОРИСТАННЯ ШУМОПОДІБНИХ КОДІВ ДЛЯ ЗАДАЧ СТЕГАНОГРАФІЇ

Розглянуто можливість використання шумоподібних кодів для задач стеганографії. Розроблено методику побудови кодових комбінацій чисел на основі теорії числових в’язанок, що дає можливість представлення кодових комбінацій чисел у вигляді шумоподібного коду для приховання інформації в найменш значущих бітах графічного формату BMP. Для цих цілей використовується технологія на основі моделі числової в'язанки, яка зводиться до заміни певних пікселів у зображенні, що дає змогу створювати ефективні алгоритми із завадостійким кодуванням та декодуванням при перетворенні графічних форматів з BMP в JPEG та інші і навпаки.

In the article the use of noise codes is examined for the tasks of steganography. The developed method of construction of code combinations of numbers is on the basis of theory of numerical bundles, which enables presentation of code combinations of numbers as a noise code for the information hiding in the the least meaningful bats of graphic format of BMP. For these aims technology is used on the basis of model of numerical bundle, which is taken to replacement of certain pels in an image, that allows to create effective algorithms with a antijammingness code and decoding at transformation of graphic formats from BMP in JPEG et al and vice versa.

Вступ

Запропоновано нетрадиційний підхід до використання шумоподібних кодів для кодування переданих даних.

Запропонований підхід відрізняється від звичайних методів кодування тим, що кодова послідовність використовується як достатньо складна кодувальна функція.