УДК 519.876.5

С.А. Панов

Формальный язык описания структуры документов и его интерпретация в формат метода компонентных цепей

Статья посвящена алфавиту и правилам формального языка L(G), используемого для описания структуры документов и интерпретации символов этого языка в компьютерные модели, создаваемые с помощью метода компонентных цепей. Разработанные компьютерные модели будут использоваться в интерактивном генераторе отчетных форм, представляющем собой систему автоматизированного документирования.

Ключевые слова: автоматизация, формальный язык, документирование, автоматизация документирования, метод компонентных цепей, компьютерное моделирование.

С целью повышения скорости создания документов и точности содержащихся в них данных активно разрабатываются и повсеместно используются системы автоматизированного документирования. К сожалению, существующие к настоящему моменту программные средства автоматизированного документирования не позволяют описывать формальную структуру документов и соответственно, направлены на создание документов только определенного формата и только для определенной области применения (промышленность, экономика, образование и т.д.). Разработка универсального формального языка позволит создать систему автоматизированного формирования документов любого состава и назначения.

В текущий момент в ТУСУРе разрабатывается собственная универсальная система автоматизированного документирования, называемая «Интерактивный генератор отчетных форм» (ИГОФ). ИГОФ предоставляет широкие возможности интерактивной работы над документом [1]. В данной работе под документом понимается набор структурных элементов (фрагментов), использующийся для хранения и передачи информации. Фрагменты могут быть двух видов: элементарные фрагменты представляют простейшие неделимые элементы, а составные фрагменты содержат в себе элементарные и другие фрагменты.

Формальный язык описания структуры документов. Модель документа в ИГОФ создается с помощью метода компонентных цепей [2], который идеально подходит для моделирования информационных процессов и систем. Сущность данного метода заключается в том, что объект (в данном случае это документ) представляет собой цепь из специальных взаимосвязанных между собой компонентов. Каждый компонент в такой цепи имеет входные и выходные узлы, с помощью которых выполняется соединение компонентов между собой. На входной узел поступают данные, которые обрабатываются и отправляются на другие компоненты с помощью выходного узла.

Задачу построения компонентов ИГОФ существенно упрощает использование формального языка описания структуры документов, каждое слово в котором отражает определенный фрагмент документа.

Формальный язык $L(G)$ задается формальной грамматикой G:

$$G = (N, T, S, P),$$

где N – набор (алфавит) нетерминальных символов (нетерминалов):

$$N = \{Документ, Текст, Таблица, Рисунок, МатВыражение, ИсЛит, Число, Слово, Цифра, ЗнакПрепинания, ЗнакОкончания, Буква, АрифмОп\},$$

- T – набор (алфавит) терминальных символов (терминалов):

$$T = \{0,1,2,3,4,5,6,7,8,9\} \cup \{\ldots,”,”,:,–\} \cup \{!,?,?!,!!!,\ldots\} \cup$$
$$\{А,Б,В,Г,Д,Е,\ldots,Я,а,б,в,г,д,е,\ldots,я,A,B,C,D,E,\ldots,Z,a,b,c,d,e,\ldots,z\} \cup \{+,–,*,/\} \cup \{□\},$$

где □ – остальные символы ASCII или Unicode; S – стартовый (начальный) нетерминал (источник):

$$S = Документ,$$

P – конечный набор (множество) правил формальной грамматики.

Таким образом, нетерминальные символы соответствуют составным, а терминальные – элементарным фрагментам документа.

Основным обязательным элементом формальной грамматики является набор специальных правил, по которым каждый нетерминальный символ может быть представлен в виде совокупности терминальных символов. Такие правила имеют вид: «левая часть» → «правая часть», где «левая часть» – это нетерминал, а «правая часть» – терминал, нетерминал или их совокупность. Так как в разрабатываемом формальном языке $L(G)$ имеется 13 нетерминалов, соответственно имеется 13 правил формальной грамматики:

1. Символ ***<Документ>*** является начальным и служит для описания всего документа в целом:

<Документ> → *<Текст>* | *<Текст><Таблица>* | *<Текст><Рисунок>* | *<Текст><Документ>*

2. Символ ***<Текст>*** представляет собой текст, который включает в себя набор всех символов из таблицы ASCII (или Unicode) и поэтому синтаксическому анализу не подлежит. В самом простом случае – тексте, где нет цифровых обозначений, рисунков, формул и ссылок на литературные источники, символ *<Текст>* может быть представлен в виде

<Текст> → *<Предложение>* | *<Предложение><Текст>*

<Предложение> → *<Слово><КонецПредложения>*

<КонецПредложения> → *<Слово><ЗнакОкончания>* | *<Слово><КонецПредложения>* | *<ЗнакПрепинания><КонецПредложения>* | *<ЗнакОкончания>*

<Слово> → *<Буква>* | *<Слово><Буква>*

3. Символ ***<Таблица>*** служит для обозначения в документе таблиц различной формы:

<Таблица> → *<Строка>* | *<Столбец>* | *<Строка><Столбец>* | *<Таблица><Строка>* | *<Таблица><Столбец>*

<Строка> → *<Ячейка>* | *<Строка><Ячейка>*

<Столбец> → *<Ячейка>* | *<Столбец><Ячейка>*

<Ячейка> → *<Текст>* | *<МатВыражение>* | *<Формула>* | *<Рисунок>* | *<Число>*

4. Символ ***<Рисунок>*** служит для обозначения рисунков с подписью:

<Рисунок> → *<Изображение>* | *<Изображение><Описание>*

Символ *<Изображение>* является терминальным и представляет собой изображение в формате JPEG, PNG, GIF, TIFF.

<Описание> → *<НачОп><Слово>* | *<НачОп><Слово><КонОп>*

<НачОп> → *Рис. <Число>*

<КонОп> → *<Слово>* | *<Слово><КонОп>* | *<ЗнакПрепинания><КонОп>* | *<ЗнакПрепинания> <Слово>*

5. Символ ***<МатВыражение>*** служит для обозначения математических выражений и формул [3].

6. Символ ***<ИстЛит>*** служит для обозначения в документе номеров источников использованной литературы:

<ИстЛит> → *<НачалоИстЛит><НомерИсточника><КонецИстЛит>*

<НачалоИстЛит> → [

<КонецИстЛит> →]

<НомерИсточника> → *1* | *2* | *3* | *...* | *N*,

где *N* – целое положительное число.

Ниже представлены символы для обозначения чисел, слов, цифр, знаков препинания, знаков окончания предложения, букв и арифметических операций:

7. ***<Число>*:**

<Число> → [–]*<Цифра>* | [–]*<Число><Цифра>*

8. ***<Слово>*:**

<Слово> → *<Буква>* | *<Слово><Буква>*

9. ***<Цифра>*:**

<Цифра> → *0* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9*

10. ***<ЗнакПрепинания>*:**

<ЗнакПрепинания> → *...* | *,* | *:* | *-*

11. ***<ЗнакОкончания>*:**

<ЗнакОкончания> → *.* | *!* | *?* | *?!* | *!!!* | *...*

12. ***<Буква>*:**

<Буква> → *А* | *Б* | *В* | *Г* | *Д* | *Е* | *...* | *Я* | *а* | *б* | *в* | *г* | *д* | *е* | *...* | *я* | *A* | *B* | *C* | *D* | *E* | *...* | *Z* | *a* | *b* | *c* | *d* | *e* | *...* | *z*

13. ***<АрифмОп>*:**

<АрифмОп> → + | – | * | /

Разработанная грамматика содержит всю полноту описания фрагментов документов различного назначения (технических, экономических, учебных, научных и т.д.).

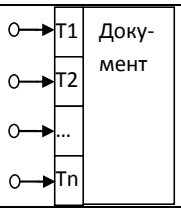
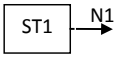
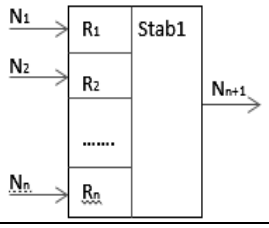
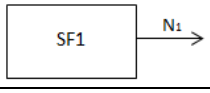
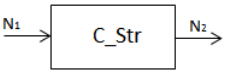
С целью разработки компьютерных имитационных моделей фрагментов документа и последующей реализации этих моделей в виде компонентов ИГОФ необходимо выполнить интерпретацию между фрагментами документа, основными нетерминальными символами в формальном языке L(G) и компонентами ИГОФ.

Интерпретация символов формального языка $L(G)$ в компьютерные имитационные модели. Интерпретация осуществляется между:

1) фрагментами документа;
2) основными нетерминальными символами формального языка $L(G)$;
3) компьютерными имитационными моделями фрагментов документа, разработанными с помощью метода компонентных цепей и реализованными в виде компонентов ИГОФ.

На текущий момент спроектировано шесть основных компонентов ИГОФ, соответствующих нетерминальным символам формального языка $L(G)$: документ, источник текста, формирователь таблицы / расширенная таблица, схема / диаграмма / график / рисунок, источник формул и числовой интерпретатор (таблица).

Компонентный базис ИГОФ

№ п/п	Фрагмент документа	Нетерминальный символ в языке формальном языке $L(G)$	Название компонента ИГОФ	Графическое представление в ИГОФ	Краткое описание
1	Документ	<Документ>	«Документ»		Служит для сбора информации с других компонентов и формирования итогового документа
2	Текст	<Текст>	«Источник текста»		Служит для вставки в документ текста
3	Таблица	<Таблица>	«Формирователь таблицы» / «Расширенная таблица»		Служит для вставки в документ таблиц [4]
4	Рисунок	<Рисунок>	«Схема», «Диаграмма», «График», «Рисунок»		Служит для вставки в документ рисунков
5	Математическое выражение	<МатВыражение>	«Источник формул»		Служит для вставки в документ формул
6	Число	<Число>	«Числовой интерпретатор»		Служит для задания числовых констант или переменных

В процессе разработки ИГОФ могут быть получены новые компоненты для работы с фрагментами документов, поэтому таблица будет постепенно пополняться все новыми и новыми элементами.

Заключение. Предложенный формальный язык $L(G)$ разработан с целью упрощения процесса создания компьютерных моделей фрагментов документов и является универсальным языком описания структуры документов различного назначения. Формальный язык $L(G)$ может быть встроен в проектируемые и уже используемые системы автоматизированного документирования, что позволит убрать ограничение таких систем на область их применения. Полнота языка подтверждается наличием всех обязательных элементов: алфавита терминальных и нетерминальных символов, грамматики и правил построения слов.

Исследование выполнено при финансовой поддержке гранта РФФИ 11-07-00384 «Метод многоуровневого моделирования алгоритмов управления технологическими процессами в сложных системах».

Литература

1. Ганджа Т.В. Задачи и архитектура подсистемы документирования исследований в среде многоуровневого моделирования МАРС / Т.В. Ганджа, С.А. Панов // Доклады Том. гос. ун-та систем управления и радиоэлектроники. – 2011. – № 2(24), ч. 2. – С. 334–338.

2. Дмитриев В.М Автоматизация моделирования промышленных роботов // В.М. Дмитриев, Л.А. Арайс, А.В. Шутенков. – М.: Машиностроение, 1995. – 304 с.

3. Дмитриев В.М. Алгоритм формирования и вычисления математических выражений методом компонентных цепей / В.М. Дмитриев, Т.В. Ганджа // Математические машины и системы. – 2010. – № 3. – С. 9–21.

4. Ганджа Т.В. Генератор табличных форм как компонент системы автоматизированного документирования / Т.В. Ганджа, С.А. Панов // Наука. Технологии. Инновации: матер. Всерос. науч. конф. молодых ученых, Новосибирск, 29 ноября – 2 декабря 2012 г.: в 7 ч. – Новосибирск: Изд-во НГТУ, 2012. – Ч. 3. – С. 88–91.

Панов Сергей Аркадьевич
Аспирант каф. моделирования и системного анализа ТУСУРа
Тел.: 8 (382-2) 41-39-15
Эл. почта: spytech@ieee.org

Panov S.A.

A formal language for describing the structure of technical documents and interpretation of it to the format method of component circuits

The paper describes a formal language that contains terminal and non-terminal symbols that reflect the structure of the technical documentation and interpretation of these symbols in the corresponding computer model presented in the format of a method of component circuits. The development of these models will create a system of automatic formation of documents which is an important task in the field of automation and process control industries.

Keywords: automation, formal language, documentation, document automation, technical paper, the method of component circuits, computer simulation.
