

Для каждого состояния A_x , $x = 1, 2, \dots, N$, проверяется условие клонирования. Клонирование состояний осуществляется следующим образом:

1. Для каждого входа состояния A_x вычисляются вероятности второго порядка (вектор вероятностей);
2. Для идентификации кластеров векторов вероятностей применяется алгоритм K -средних:
 - 2.1. K – количество кластеров (каждый кластер соответствует новому состоянию). Полагается $K=2$.
 - 2.2. Векторы распределяются между кластерами случайным образом в соответствии с равномерным распределением.
 - 2.3. Вычисляются центроиды кластеров, как вероятности перехода, полученные после перераспределения весов входных связей, включенных в данный кластер.
 - 2.4. Для каждого входа выполняется: если ближайший центроид относится к другому кластеру, то необходимо отнести данный вход к этому кластеру. Переход на 2.3, до тех пор, пока есть перемещения между кластерами.
 - 2.5. Проверка, является ли данное решение точным. Решение называется «точным», если для каждого кластера выполняется условие, что вероятности, заданные его центроидом, близки к вероятностям второго порядка всех входов, включенных в данный кластер, т.е. не существует входа, для которого разность между вероятностью второго порядка и вероятностью соответствующего центроида превышает заданный предел γ .
 - 2.6. Если решение не является точным, выполнить алгоритм K -средних еще раз, про $K=K+1$ (при $K \leq I$, для $K = I$ решение однозначно). При реализации алго-

ритма, K можно увеличивать в геометрической прогрессии $K = K^2$, иначе для состояний с большим количеством входов большое количество итераций алгоритма кластеризации приведет к недопустимому времени выполнения.

3. Создать $K-1$ клонов (копий) состояния A_x и перераспределить входные связи между клонами в соответствии с результатом кластеризации. Вычислить веса выходов.

Результаты вычислительных экспериментов на реальных и сгенерированных данных показали, что время работы алгоритма линейно зависит от сложности модели, а полученное количество состояний данной модели меньше, чем в моделях N -грамм эквивалентной адекватности этих моделей.

Литература

1. *Borges J., Levene M.* (2000). Data mining of user navigation patterns // *Masand B. and Spiliopoulou M.* (editors). Usage Analysis and User Profiling, Lecture Notes in Artificial Intelligence (LNAI 1836), P. 92–111. Springer-Verlag, Berlin.
2. *Cadez I., Heckerman D., Meek C., Smyth P., White S.* (2000). Visualization of navigation patterns on a web site using model based clustering // *Proceedings of the International KDD conference*. 2000. P. 280–284.
3. *Zhu J., Hong J., Hughes J.G.* Using markov models for web site link prediction // *Proceedings of the 13th Annual Conference on Hypertext and Hypermedia*. 2002. P. 169–177.

Южно-Российский государственный технический университет
(Новочеркасский политехнический институт)

10 июля 2007 г.

УДК 004.91

ПОДХОД К ПОСТРОЕНИЮ СПЕЦИАЛИЗИРОВАННЫХ ЭЛЕКТРОННЫХ БИБЛИОТЕК

© 2007 г. А.Н. Иванченко, М.В. Кислов

Введение

На сегодняшний день в различных сферах деятельности человека накоплены большие объемы различного рода справочной информации, и процесс её накопления неуклонно растёт.

Ввиду глобальной информатизации общества все актуальнее становится задача перевода в цифровую форму хранения уже накопленных информационных массивов. Перевод документов из бумажной формы хранения в цифровую дает возможность применения современных информационных технологий, что позволяет более эффективно использовать имеющиеся

информационные массивы. Создание и использование электронных библиотек (ЭБ) уже доказало свою эффективность, а использование электронных документов оказалось намного удобнее их бумажных аналогов.

Однако существуют важные аспекты при выборе технологии перевода бумажных документов и способа хранения их электронных версий. Выбор технологии перевода документов часто зависит от технологии последующего использования. Самая распространенная на сегодняшний день технология использования документов – это библиотечная. Способ хранения документов при данной технологии характеризуется

что существует сам документ и его метаданные. Основной формой хранения информации считается документ. Альтернативной можно считать «технология использования структурированных документов». В этом случае документ хранится разделенный на некоторые порции, что позволяет использовать документ не только как единое целое, но и работать с отдельными его фрагментами.

Преимуществом такого способа хранения является то, что документ может представляться в виде многоуровневой древовидной структуры, а это в свою очередь дает возможность эффективнее осуществлять навигацию по документу. В пользу данного способа хранения может выступать то, что документы изначально создаются в структурированном виде. Практически каждый документ имеет оглавление, которое фактически является его древовидной структурой.

Такой способ хранения документов применяется в различных системах обучения, так как позволяет обращаться к определенным частям документа. Примерами таких систем могут являться АСОП-«Эксперт» [1], АСОП-«Наставник» [2], Центурион, Эксперт и другие им подобные [3]. Также необходимо отметить, что данный способ хранения документов, разделенных на порции, можно рассматривать как реализацию «библиотечной» технологии использования.

На сегодня системы обучения и проверки знаний находят широкое применение в различных отраслях [4 – 8]. В связи с этим актуальной становится задача создания и ведения полнотекстовых ЭБ для таких специализированных систем. Объем документов в ЭБ может достигать сотен и тысяч экземпляров, поэтому процессу автоматизации перевода документов в структурированный вид хранения необходимо уделять должное внимание.

Существуют различные методы и средства преобразования текста документа из его линейного вида в древовидную структуру. Оценкой эффективности данных систем может являться не только качество результата перевода, но и требования к предварительной подготовке текста исходных документов. Например, для добавления документа в ЭБ программного комплекса АСОП-«Эксперт» текст исходного документа должен быть размечен специальными системными метками, а в программном комплексе АСОП-«Наставник» накладываются жесткие ограничения на стилевое оформление документа. Это негативным образом сказывается на эффективности применения подобных ЭБ, так как предъявляются дополнительные требования к квалификации людей выполняющих эти задачи.

В данной статье рассматривается подход к созданию инструмента, позволяющего формировать ЭБ, используя специальный формат хранения, который позволяет адекватно отображать текст документов в виде древовидной структуры. Описывается один из способов реализации метода обработки документов, позволяющий автоматически восстанавливать древовидную структуру документа по его тексту.

С помощью данного метода происходит преобразование текста из линейного вида в специальный формат хранения.

Отличительной чертой ЭБ, используемых в системах обучения, от ЭБ в их «классическом понимании» являются сами документы. Основной составляющей таких библиотек являются нормативно-технические документы (НТД), которые представляют собой специализированные отраслевые документы, нормативные акты, федеральные законы, ГОСТы и т.д. Особенности их заключаются в способе изложения информации. Как правило, тексты НТД сильно структурированы и представляют собой набор параграфов, сгруппированных по разделам, главам, частям и т.д. (рис. 1).

Правила технической эксплуатации электрических станций (утв. приказом Минэнерго РФ от 19	
1. Организация эксплу	
1.1. Основные положения	
1.1.1.	Настоящие Правила распространяются на тепл органическом топливе, гидроэлектростанции, электрические на организации, выполняющие работы применительно к этим
1.1.2.	При выполнении на ГЭС АСУ ТП должны выполняться
1.1.3.	Опытные (экспериментальные), опытно-промышлен подлежат приемке в эксплуатацию приемочной комиссией, ее или выпуску продукции, предусмотренной проектом.
1.2. Персонал	
1.2.1.	К работе на энергообъектах электроэнергетики образованием, а по управлению энергоустановками также и с
1.2.2.	Лица, не имеющие соответствующего профессионала вновь принятые, так и переводимые на новую должность д отрасли форме обучения.

Рис. 1. Пример текста документа

Принимая во внимание специфичность применения ЭБ НТД, вопросы, касающиеся полнотекстовых ЭБ литературных изданий, в данной статье не рассматриваются.

Основные принципы структуризации информации в нормативно-технических документах

Основываясь на том, что текст НТД разделен на порции (параграфы), введем понятие «элементарного объекта», который представляет собой определенную порцию текста и на основе которого в дальнейшем будем производить построение древовидной структуры документа.

В процессе анализа текстов НТД было установлено, что элементарный объект может содержать следующие элементы текста: название документа, раздела, подраздела, части, главы и т.д., а также их номера, номера параграфов и текст параграфов и разделов. Некоторые из приведенных элементов могут быть сходны по своей структуре построения, поэтому будет целесообразно сгруппировать их по определенным признакам:

- идентификатор объекта;
- название объекта;
- текст объекта.

Данные группы элементов будем считать подобъектами, составляющими элементарный объект.

Учитывая принципы построения текста, подобъекты располагают в определенной последовательности относительно друг друга. Например, мы знаем, что номер главы не может располагаться после её названия или что вслед за номером параграфа обязательно должен располагаться его текст. В процессе анализа текстов НТД было установлено несколько вариантов комбинации подобъектов в элементарном объекте. Их можно описать следующей синтаксической диаграммой на рис. 2. Примеры вариантов комбинаций подобъектов приведены в табл. 1.

Будем считать, что элементарный объект может быть однозначно определен в тексте документа либо идентификатором объекта, либо названием объекта, либо их парой. Для упрощения последующей формализации описания структуры документа введем дополнительный термин «искусственный идентификатор». Считаем, что «искусственным идентификатором» будет определяться элементарный объект, не содержащий идентификатор объекта. Фактически «искусственный идентификатор» не содержит ничего, он только обозначает место, где должен был бы находиться идентификатор объекта.

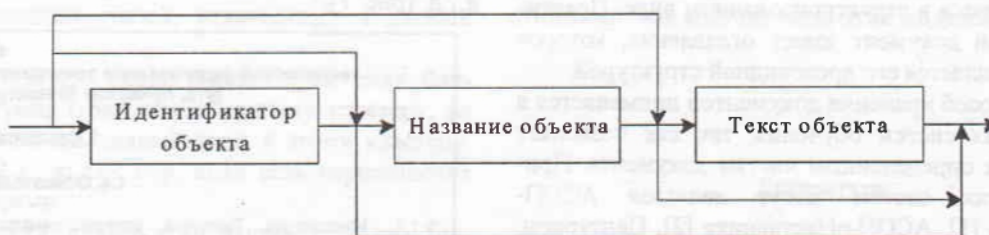


Рис. 2. Синтаксическая диаграмма элементарного объекта

Таблица 1

Примеры вариантов комбинаций подобъектов

Комбинация следования подобъектов	Пример
Идентификатор объекта	Примером элементарного объекта, который содержит только идентификатор, может быть текст: «ЧАСТЬ ПЕРВАЯ», «Часть вторая» и т.д.
Идентификатор объекта → Название объекта	Примером данной комбинации подобъектов может быть текст вида: «Раздел А.»+«Название», «Глава 1.»+«Название» и т.д.
Идентификатор объекта → Текст объекта	Третий вариант наиболее распространен в текстах документов и ему соответствует «лист» в древовидной структуре документа. Примером данной последовательности подобъектов является текст вида: «1.1.3.»+«Текст параграфа», «Статья 12.»+«Текст статьи» и т.д.
Идентификатор объекта → Название объекта → Текст объекта	Данный пример последовательности подобъектов встречается довольно редко. Примером может являться фрагмент, в котором между названием раздела и названием подраздела присутствует вступительный текст: «Номер»+«Название раздела»+«Текст»+далее идут подразделы. Или может встречаться другой случай: «Номер»+«Название подраздела»+«Текст»+далее идут параграфы
Название объекта	Данный пример элементарного объекта может встречаться, когда в виде «Названия объекта» выступает «Название документа» или название группы параграфов в подразделе
Название объекта → Текст объекта	Примером данной комбинации подобъектов может быть текст вида: «ВВЕДЕНИЕ»+«Текст введения», «ЗАКЛЮЧЕНИЕ»+«Текст заключения» и т.д.
Текст объекта	Пример элементарного объекта такого вида практически не встречается так как текст, как правило, относится к разделу, подразделу или параграфу

Математический аппарат

Для реализации метода преобразования линейного текста документа в специальный вид предлагается использовать математический аппарат теории формальных языков. Данный метод предполагается разбить на следующие этапы:

1. Определение элементарных объектов в тексте документа.
 2. Определение иерархических связей (древовидной структуры) полученных элементарных объектов.
- В общем виде поэтапную обработку текста можно представить, как показано на рис. 3.

Исходные данные	Этап 1	Этап 2
Текст на естественном языке	Текст на промежуточном специализированном языке	Древовидная структура
<p>1. Организация эксплуатации</p> <p>1.1. Основные положения и задачи</p> <p>1.1.1. Настоящие Правила распространяются на тепловые электростанции, работающие на органическом топливе...</p> <p>1.1.2. При выполнении на ГЭС АСУ ТП должны выполняться положения раздела 4.7 настоящих Правил...</p>	<pre> <object id=001 type=1>1.</object> <header id=001>Организация эксплуатации</header> <object id=002 type=2>1.1.</object> <header id=001>Основные положения и задачи</header> <object id=003 type=3>1.1.1.</object> <text id=003>Настоящие Правила распространяются на тепловые электростанции, работающие на органическом топливе...</text> <object id=004 type=3>1.1.2.</object> <text id=003>При выполнении на ГЭС АСУ ТП должны выполняться положения раздела 4.7 настоящих Правил...</text> </pre>	

Рис. 3. Пример поэтапной обработки текста

Этап определения элементарных объектов связан с определением в тексте всех возможных идентификаторов объектов, так как один элементарный объект может быть однозначно определен одним идентификатором.

В связи с тем что текст НТД имеет сложный вид, могут возникать ситуации, когда в тексте документа встречаются нумерованные списки, ссылки на другие параграфы и другие элементы, представляющие собой различные комбинации цифр. При определении элементарных объектов такие элементы могут быть определены как «идентификаторы объектов», что влечет за собой неправильную интерпретацию иерархической структуры документа. Здесь возникает дополнительная задача отсеивания ложных идентификаторов объектов.

Решение задачи определения в тексте элементарных объектов построим с использованием математического аппарата теории формальных языков. Для текста документа представим в виде цепочки

символов, сформированной по правилам формального языка, адекватно описывающего структуру НТД. К таким правилам можно отнести правила построения идентификатора объекта (номер главы, раздела, параграфа), названия объекта и его текста. Пример правила для формирования номера параграфа показан на рис. 4.

Задачу определения элементарных объектов предполагается выполнять в два этапа:

1. Выделение всех идентификаторов объектов.
2. Отсеивание ложных идентификаторов объектов.

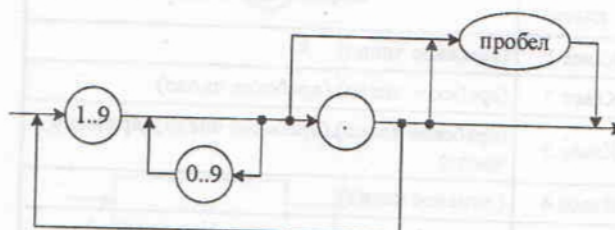


Рис. 4. Синтаксическая диаграмма формирования номера параграфа

Первый этап требует проведения лексического анализа текста, в задачи которого входит нахождение во входной цепочке последовательности символов, образующих отдельные лексические элементы, называемые лексемами. В нашем случае лексемами будут являться идентификаторы объектов. Для выполнения данного этапа необходимо построить группу правил, с помощью которых во входной цепочке (тексте документа) будут определяться все необходимые виды лексем.

Второй этап требует проведения синтаксического и семантического анализа. Задача семантического анализа предполагает построение системы правил для анализа лексем. Одновременно приходится анализировать контекст (семантику) нескольких лексем (идентификаторов объектов), так как их расположение относительно друг друга имеет некоторые ограничения.

Необходимость семантического анализа обусловлена наличием ситуаций, при которых некоторые последовательности символов могут быть ошибочно интерпретированы как идентификаторы объектов. Примером могут служить ссылки в тексте элементарного объекта на другие элементарные объекты (параграфы), нумерованные списки и т.д.

Анализ элементарных объектов

Анализ коллекций НТД показал, что для различных видов документов используются определенные варианты формы записи идентификаторов объектов. В результате анализа формы записи идентификаторов были выделены определенные классы, приведенные в табл. 2. Для представления формы подобъектов будем использовать нотацию языка расширенных регулярных выражений. Введем следующие обозначения:

- арабское число – число, записанное в арабской системе исчисления;
- римское число – число, записанное в римской системе исчисления;
- буква – символ из русского алфавита;
- «...» – обозначение последовательности символов;
- (...) – служит для наглядного отображения группы символов;
- | – обозначение альтернативы.

Таблица 2

Варианты представления подобъекта «Идентификатор объекта»

Название класса	Форма подобъекта
Класс 1	(арабское число)
Класс 2	(арабское число).(арабское число)
Класс 3	(арабское число).(арабское число).(арабское число)
Класс 4	(римское число)
Класс 5	«Часть» («первая» «вторая» «третья» ...)
Класс 6	«Раздел» ((буква) (арабское число) (римское число))
Класс 7	«Глава» ((арабское число) (римское число))
Класс 8	«Статья» (арабское число)
Класс 9	«Статья» (арабское число).(арабское число)
Класс 10	Искусственный идентификатор для основных названий
Класс 11	Искусственный идентификатор для основных приложений
Класс 12	Искусственный идентификатор для обычных названий

Возможные варианты форм представления «названия объекта», которые могут не иметь идентификаторов, но могут определять элементарный объект и участвовать в иерархии документа, приведены в табл. 3.

Таблица 3
Варианты представления подобъекта «Название объекта»

Название класса	Форма подобъекта
Основные названия	«Оглавление», «Содержание», «Предисловие», «Введение», «Общая часть», «Заключение», «Список литературы», «Список использованной литературы», «Литература»
Приложения	«Приложение» ((буква) (арабское число))
Обычные названия	Обычный текст, содержащий алфавитно-цифровые символы и символы перехода на новую строку

Принадлежность символьной цепочки к классам, приведенным в табл. 3, можно определить не только по её контексту, но и по некоторым стилистическим свойствам. Приведем список таких свойств в табл. 4.

Таблица 4
Свойства текста, принадлежащего к классам «Основные названия», «Приложения», «Обычные названия»

№	Описание
1	Все буквы в верхнем регистре
2	Текст выровнен по центру
3	Текст выровнен по правому краю
4	Текст выделен жирным шрифтом
5	Текст находится между двумя подряд идущими спецсимволами «перевод строки» слева и двумя спецсимволами «перевод строки» справа

Для унификации всех комбинаций элементарных объектов используем искусственные идентификаторы, которыми будут обозначаться элементарные объекты, не имеющие собственных идентификаторов (объекты, принадлежащие классам, приведенным в табл. 3).

Формальное описание классов подобъектов

Для формального описания классов подобъектов целесообразно будет использовать математический аппарат теории формальных языков. В данной работе нами будет разработан формальный язык, адекватно описывающий структуру нормативно-технического документа. Для описания языка будем использовать грамматику и синтаксические диаграммы. Синтаксические диаграммы являются эквивалентным представлением грамматики языка, но имеют преимущество для визуального представления в силу своей наглядности (рис. 5).

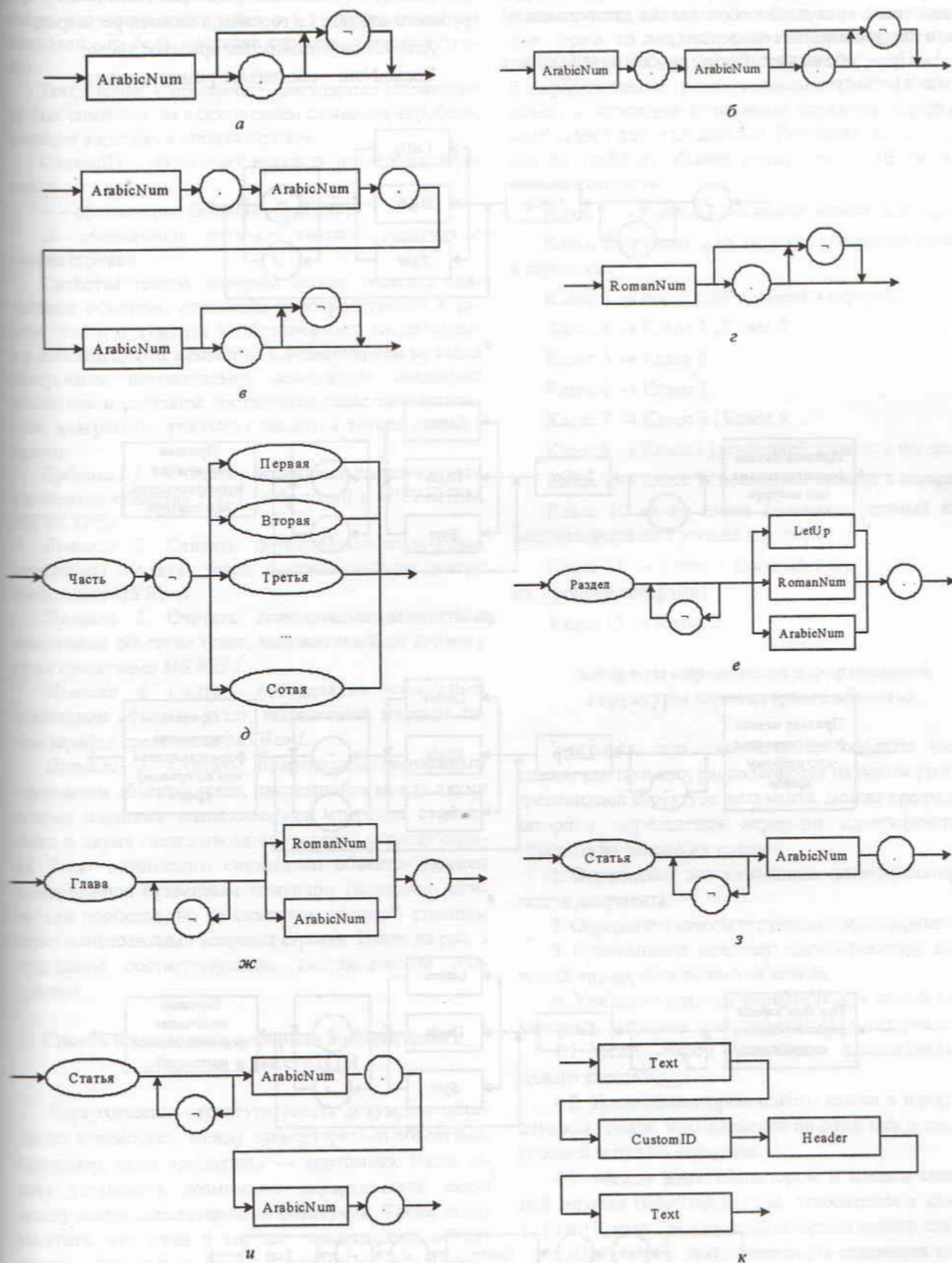


Рис. 5. Синтаксические диаграммы описания идентификаторов объектов: а – для класса 1; б – для класса 2; в – для класса 3; г – для класса 4; д – для класса 5; е – для класса 6; ж – для класса 7; з – для класса 8; и – для класса 9; к – для классов 10, 11, 12

При составлении синтаксических диаграмм будем использовать следующие обозначения для терминальных и нетерминальных символов (рис. 6):

LetUp – обозначает букву русского алфавита в верхнем регистре;

Letters – обозначает регулярное множество букв русского алфавита в верхнем и нижнем регистрах;
 ArabicNum – обозначает арабское число;
 RomanNum – обозначает римское число;

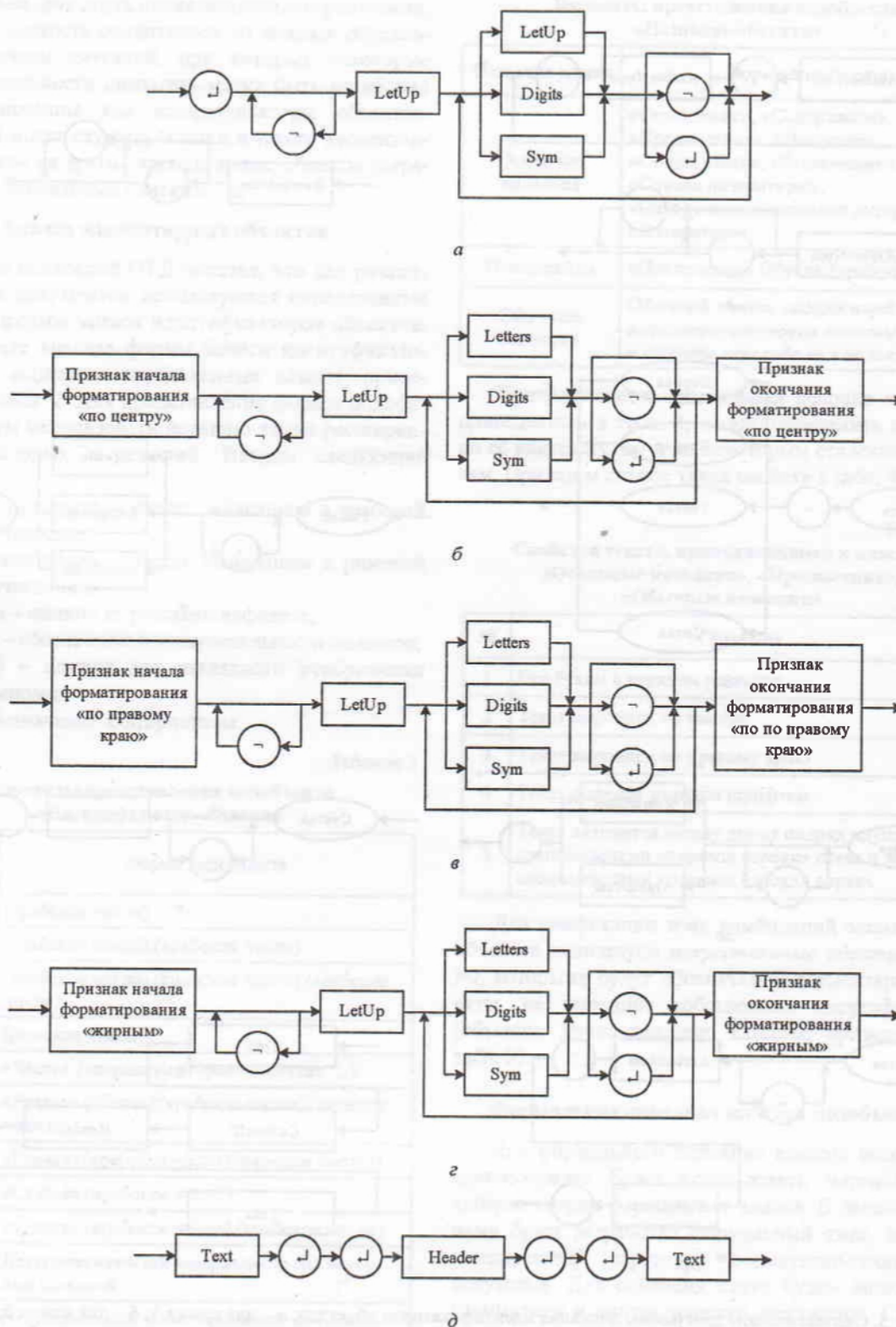


Рис. 6. Синтаксические диаграммы определения границ заголовков: а – для правила 1; б – для правила 2; в – для правила 3; г – для правила 4; д – для правила 5

— обозначает множество символов, например знаки препинания и т.д., за исключением «пробел», «возврат каретки» и «новая строка»;

Header — обозначают регулярное множество символов, за исключением символов «пробел», «пробел» и «новая строка»;

CustomID — обозначает искусственный идентификатор;

— обозначение символа «пробел»;

— обозначение символов «возврат каретки» и «новая строка».

Свойства текста, который может являться «названием объекта», отражены непосредственно в самом тексте документа и обозначаются специальными комбинациями символов. Сформулируем правила построения потенциально возможных «названий объектов» и составим соответствующие синтаксические диаграммы, учитывая свойства текста данного документа.

Правило 1. Считать потенциально-возможным «названием объекта» текст, набранный в «ВЕРХНЕМ РЕГИСТРЕ».

Правило 2. Считать потенциально-возможным «названием объекта» текст, выровненный по центру средствами MS Word.

Правило 3. Считать потенциально-возможным «названием объекта» текст, выровненный по правому краю средствами MS Word.

Правило 4. Считать потенциально-возможным «названием объекта» текст, выделенный жирным типом шрифта средствами MS Word.

Правило 5. Считать потенциально-возможным «названием объекта» текст, находящийся между двумя идущими спецсимволами «перевод строки».

и двумя спецсимволами «перевод строки» справа и слева.

Текст возможного «названием объекта» должен начинаться буквенным символом (возможно символом пробелов, но не символом «точка») стоящим перед спецсимволами «перевод строки». Выше на рис. 5 показаны соответствующие синтаксические диаграммы.

и спецсимволами «перевод строки».

Способ определения иерархии элементарных объектов в тексте НТД

Иерархическая структура текста документа определяется взаимосвязи между элементарными объектами. Например, связь «родитель» → «потомок». Наша задача — установить возможные иерархические связи между всеми элементарными объектами. Также стоит отметить, что один и тот же элементарный объект может являться потомком для одного объекта и родителем для другого, а также не иметь потомков вообще.

Материал в документах излагается последовательно — иерархия не нарушается, следовательно, элементарные объекты также будут встречаться последовательно и элементарный объект, встретившийся в тексте раньше, в иерархии будет находиться выше

при выполнении некоторых условий. Учитывая, что элементарный объект может быть определен идентификатором, то для построения иерархии идентификаторов можно использовать их классы. В соответствии с определенными ранее классами идентификаторов объектов приведем возможные варианты иерархических связей для этих классов. Результат сведем в список по шаблону «Класс родителя» → «Возможные классы потомка»:

Класс 1 → Класс 2 | последний элемент в иерархии.

Класс 2 → Класс 3 | Класс 12 | последний элемент в иерархии.

Класс 3 → последний элемент в иерархии.

Класс 4 → Класс 1 | Класс 2.

Класс 5 → Класс 6.

Класс 6 → Класс 7.

Класс 7 → Класс 8 | Класс 9.

Класс 8 → Класс 1 | последний элемент в иерархии.

Класс 9 → Класс 1 | последний элемент в иерархии.

Класс 10 → не имеет потомков {данный класс располагается на 1 уровне иерархии}.

Класс 11 → Класс 1 {данный класс располагается на 1 уровне иерархии}.

Класс 12 → Класс 3.

Алгоритм определения иерархической структуры элементарных объектов

Учитывая, что элементарные объекты одного класса, как правило, располагаются на одном уровне в древовидной структуре документа, можно предложить алгоритм определения иерархии идентификаторов объектов на основе их классов:

1. Определяем все возможные идентификаторы в тексте документа.

2. Определяем классы полученных идентификаторов.

3. Присваиваем каждому идентификатору порядковый номер, независимо от класса.

4. Учитывая правила иерархических связей между классами, выбираем идентификаторы в интервале:

4.1. Между парой ближайших идентификаторов одного класса.

4.2. Идентификатором одного класса и идентификатором класса, находящегося на один или несколько уровней выше по иерархии.

4.3. Между идентификатором и концом символьной цепочки (идентификаторы, относящиеся к классам: 1, 10 и 11, считаем идентификаторами одного класса).

5. Выбранные идентификаторы считаются потомками первого идентификатора.

6. Пункты 4 и 5 выполняются последовательно для всех идентификаторов в соответствии с правилами наследования.

Результат работы алгоритма, схематично, можно отобразить на рис. 7.

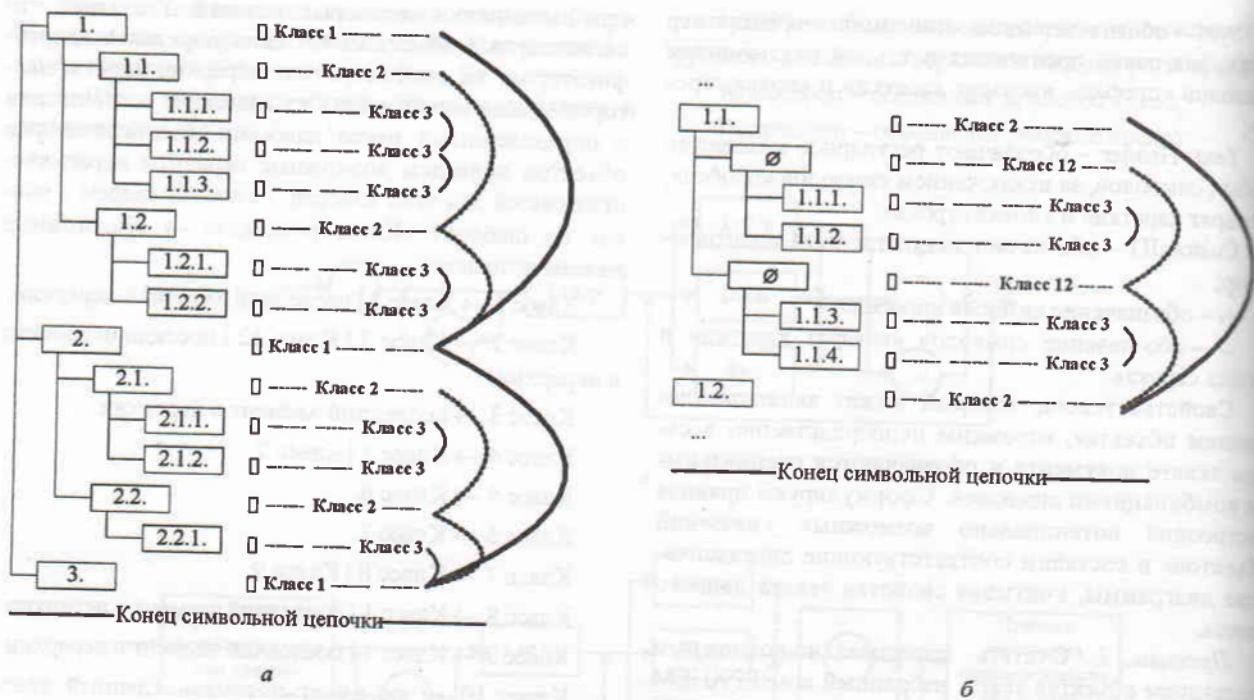


Рис. 7. Пример иерархической структуры документа

Заключение

В статье предлагается метод обработки документов, позволяющий автоматически восстанавливать древовидную структуру НТД по его тексту. Данный метод можно использовать при создании специализированных ЭБ, использующихся в различных системах обучения.

При использовании предложенной методики для наполнения ЭБ новыми документами от персонала не требуется высокой квалификации, а также не предъявляются жесткие требования к оформлению исходных текстов вносимых документов.

На основе данного метода ведется разработка инструмента, позволяющего в автоматизированном (полуавтоматическом) режиме добавлять документы в ЭБ НТД.

Литература

1. Информационно-образовательная среда для обучения и поддержки квалификации персонала АСОП «Инфосреда» // <http://ucpk.ru/cgi-bin/page.pl?article=67>

2. Автоматизированная система обучения персонала АСОП «Наставнику» // <http://do.ucpk.ru>.

3. Gerald Salton, James Allan, Amit Singhal. Automatic text decomposition and structuring // *Information Processing & Management*. 1996. 32(2):127-138.

4. Прокофьева Н.О. Вопросы организации компьютерного контроля знаний // *Educational Technology & Society* 2006. 9(1): 433-440.

5. Зайцева Л.В., Прокофьева Н.О. Модели и методы аддитивного контроля знаний // *Educational Technology & Society* 2004. 7(4): 265-277.

6. Добрынин В.Ю., Клюев В.В., Некрестьянов И.С. Оценка тематического подобия текстовых документов // Тр. второй Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Протвино, Сентябрь 2000. С. 204-210.

7. Барашев Д.В., Горшкова Е.А., Новиков Б.А. Оптимизация представления XML документов в реляционной базе данных // Тр. второй Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Протвино, Сентябрь 2000. С. 224-229.

8. Столяров С.В., Ермаков Н.Б., Федотов А.М. Электронные Библиотеки: схемы данных и объектная модель документа. http://www.nsc.ru/ws/show_abstract.dhtml?ru+127+9299.