ПОИСК СПОСОБОВ СНИЖЕНИЯ НАГРУЗКИ НА СИСТЕМЫ ХРАНЕНИЯ ПРЕДПРИЯТИЙ МИКРОЭЛЕКТРОННОЙ ПРОМЫШЛЕННОСТИ

Черников Б.В.[1,2], Кремер Е.А.[3], Борисова Е.А.[2]

[1]ООО «Газпром ВНИИГАЗ», г. Москва

[2]Российский экономический университет им. Г.В. Плеханова, г. Москва,

[3]Московский институт электронной техники, г. Москва

bor-cher@yandex.ru, kremerea@gmail.com, lb20062006@yandex.ru

Аннотация: Увеличение количества текстовых документов, применяемых на предприятиях, влечет за собой рост трудозатрат, направленных на создание этих документов. Применение лексикологического синтеза может позволить исправить сложившуюся ситуацию. В работе рассмотрена методика сокращения объема текстовых документов на предприятиях микроэлектронной промышленности.

Ключевые слова: лексикологический синтез, методика, хранение информации, слабоформализуемый документ, индексная последовательность.

Введение

В настоящее время документооборот на предприятиях в большинстве случаев осуществляется в бумажном виде, несмотря на широкое распространение персональных компьютеров и применение на предприятиях систем электронного документооборота. Данный факт обуславливает необходимость наличия архива бумажных документов значительных объемов.

Преобладание бумажных документов над электронными копиями определяется рядом причин, к которым можно отнести следующие:

- требования документов, регламентирующих документооборот внутри предприятия (помимо внутренних стандартов предприятия сюда входят государственные и ведомственные стандарты);
- недостаточное распространение электронных средств защиты текстовых документов, таких как электронная подпись;
- требования предоставлениям документов регулирующим органам, вышестоящим организациям и партнерам в бумажном виде;
- высокая юридическая значимость.

Документы, используемые для информационного обмена на предприятиях микроэлектронной промышленности, в подавляющем большинстве являются слабоформализуемыми, поскольку их содержание сильно зависит от конкретной ситуации, однако при этом необходимо обеспечить возможность учета всех нюансов документируемой ситуации.

Слабоформализуемые документы – полнотекстовые, табличные либо смешанные документы, содержание которых существенным образом связано с произвольной, меняющейся от конкретной ситуации структурой. Это документы, обладающие достаточно высокой степенью вариативности. В связи с этим содержательная структуризация слабоформализуемых документов может требовать детализации как взаимосвязи, так и взаимной зависимости композиции текста вплоть до атомарных значений – фрагментов фраз, слов, и даже частей отдельных слов [1].

Актуальность исследований в области документооборота, документирования, хранения и передачи информации обусловливается приоритетной программой правительства «Информационное общество» [2], а также общим увеличением количества и объема хранимых электронных документов.

Целью данного исследования является разработка методики сокращения размера слабоформализуемых документов при хранении на предприятиях микроэлектронной промышленности при помощи лексикологического синтеза.

1 Лексикологический синтез

Лексикологический синтез – формирование текстовых фрагментов путем создания фраз на основе набора опорных слов, который формируется по результатам глубокого анализа текста документа путем связывания текстовых фрагментов с конкретным опорным словом, входящих в состав фраз или выражений формируемого текста.

Основой лексикологического синтеза является тот факт, что каждая область и сфера работы на предприятии сопровождается конкретным комплексом документов. Любой документ, описывающий ситуацию в исследуемой отрасли, содержит переменную и постоянную информацию. При анализе текстового документа можно выделить постоянную информацию, характерную именно для данного вида документов. К постоянной информации добавляется переменная информация, которая может принадлежать конечному множеству вариантов, если текст заранее унифицирован. Поскольку множество вариантов конечно, то, объединив их, можно информацию отнести к разряду переменной унифицированной. При создании документа постоянная информация вносится автоматически, а переменная унифицированная информация внедряется после выбора нужных вариантов из сохраненного множества.

2 Методика снижения нагрузки на системы хранения предприятия

Рассмотрим методику сокращения объема документов с помощью лексикологического синтеза (рис. 1).

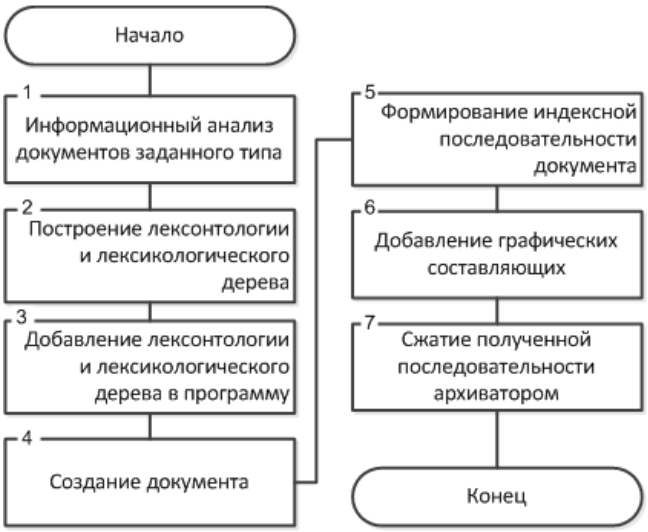

Рис. 1. Методика сокращения объема документов

1. На первом этапе необходимо провести глубокий информационный анализ того вида документов [3], создание которых планируется автоматизировать. Количество и наличие информации в документе изменяются. На присутствие или отсутствие информации в документе влияют различные факторы. Ситуаций, в которых создаваемые документы будут совпадать, практически не бывает.

Для возможности применения лексикологического синтеза необходимо проанализировать структуру информации, которая содержится в документе. Вся информация традиционно делится на постоянную и переменную. Постоянная информация – неизменная информация, которая используется в течение длительного периода времени без каких либо изменений. Переменная информация отражает фактические количественные и качественные характеристики деятельности предприятия, которые необходимо закрепить в документе. К переменной информации можно отнести всю информацию, которую необходимо вводить в документ при каждом его заполнении.

2. На втором этапе строится лексикологическая схема (лексонтология) и лексикологическое дерево документа заданного вида.

На основе лексического анализа выделяются фрагменты документа, которые определяются конкретным словом либо словосочетанием, обозначающим наличие конкретной формулировки в документе. Такие слова называются опорными. Основным критерием включением какого-либо слова или словосочетания во множество опорных слов является его однозначное понимание и применение.

Далее сформированное множество опорных слов применяется для создания лексикологической схемы (лексонтологии) и дерева документов. Лексонтология документа представляет собой модель взаимной связи опорных слов, входящих в состав сформированного множества и используемых в процессе создания документов данного вида с учетом вариативности отдельных экземпляров [1].

3. На третьем этапе сформированная лексонтология и лексикологическое дерево импортируются в исполняемую программу, с которой будут работать составители и создатели документов.

4. На четвертом этапе происходит создание документа.

5. На пятом этапе формируется индексная последовательность на основе внесенной в документ информации.

Фиксация индексной последовательности, соответствующей выбираемым опорным словам, осуществляется пошагово в рамках организуемого цикла выбора опорных слов. В случае отсутствия в лексонтологии унифицированного варианта формулировки, определяемого опорным словом, в индексную последовательность внедряется вводимый неунифицированный фрагмент. По завершении процесса формирования документа в разделе подписей фиксируется индекс подписи должностного лица (исполнителя документа), который также конкатенируется в индексный информационный пакет.

Принципиальным отличием данного варианта от классического способа формирования индексной последовательности является использование двоичных кодов, где «0» – опорное слово не выбрано, а «1» – опорное слово выбрано. В рамках каждого уровня фиксируется своя двоичная последовательность, которая переводится в десятичную систему, получая тем самым одно десятичное число на уровень. Все уровни располагаются друг за другом через пробел. В случае наличия неунифицированных фрагментов они располагаются после десятичного числа соответствующего уровня и располагаются внутри символов «(» и «)», а после завершения формирования последовательности конкатенируются в одну строку, после чего сжимаются с помощью выбранного алгоритма архивации, а потом добавляются в конец индексной последовательности, а все предыдущие неунифицированные фрагменты удаляются, и заменяются номерами, соответствующими номерам в конкатенированной строке.

6. На шестом этапе к сформированной индексной последовательности добавляются графические составляющие, такие как логотипы, картинки и т.п. В конец последовательности дописывается индекс графической составляющей.

7. На седьмом этапе полученная последовательность дополнительно сжимается архиватором.

3 Выводы

В результате исследования получены следующие результаты:

1. Предложена методика сокращения текстовых документов и доказана ее эффективность на примере технологической инструкции.

2. Разработанная методика определяет четкую последовательность действий для сокращения объема текстового документа и позволяет применять ее на предприятиях микроэлектронной промышленности.

3. В результате исследования выявлены следующие проблемы, на которые будет направлено дальнейшее исследование:

- возможен большой объем неунифицированной информации в индексной последовательности, что существенно увеличит ее размер – следует проанализировать способы сокращения ее объемов;
- необходимость возможности внедрения в документ неиндексированных графических материалов в последовательности, что влечет необходимость совершенствования метода добавления индекса графических компонентов в документ.

Литература

1. Черников Б.В. Лексикологический синтез документов в комплексах информационных систем. – М.: ИД «ФОРУМ». – 2017. – 336 с.

2. Государственная программа Российской Федерации «Информационное общество (2011-2020 годы)» (утверждена постановлением правительства Российской Федерации от 15.04.2014 №313) // Собрание законодательства РФ, 05.05.2014, № 18 (2 ч.), ст. 2159.

3. Черников Б.В., Кремер Е.А. Анализ информации текстовых документов предприятий микроэлектронной промышленности// Современные наукоемкие технологии. – 2018. – № 5, с. 168-172.