УДК 004.91:005.92

СОКРАЩЕНИЕ ОБЪЕМА ТЕКСТОВЫХ ДОКУМЕНТОВ ПРИ ХРАНЕНИИ НА ПРЕДПРИЯТИЯХ МИКРОЭЛЕКТРОННОЙ ПРОМЫШЛЕННОСТИ^{1,3}Черников Б.В., ²Кремер Е.А., ³Борисова Е.А.¹ООО «Газпром ВНИИГАЗ», Москва, e-mail: bor-cher@yandex.ru;²Московский институт электронной техники, Зеленоград, e-mail: kremerea@gmail.com;³Российский экономический университет им. Г.В. Плеханова, Москва, e-mail: lb20062006@yandex.ru

Ежедневно на предприятиях происходит множество событий и процессов, которые требуют отражения в различных документах. Во время производственного процесса необходимо фиксировать практически каждое действие. Большой объем таких действий приводит к постоянному увеличению количества создаваемых документов и, как следствие, росту объема хранилищ и баз данных, выделяемых под эти документы. Всё это приводит к повышенной нагрузке на системы хранения данных, применяемые на предприятии. Увеличение количества документов влечет увеличение трудозатрат людей, которые создают эти документы, используя традиционные способы формирования документов, когда вся текстовая информация вводится с клавиатуры. Существующие способы автоматизации документооборота и документирования не оказывают существенного влияния на сложившуюся ситуацию, а значит, не являются достаточно эффективными. Данный факт позволяет объяснить применение лексикологического синтеза, который позволит исправить сложившееся положение. Лексикологический синтез в аспекте применения его на предприятиях микроэлектроники имеет ряд особенностей. В работе рассмотрена методика сокращения объема текстовых документов, создание которой дает возможность сократить и без того ощутимые результаты в области сокращения объема текстовых документов при хранении.

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

REDUCTION OF TEXT DOCUMENTS VOLUME DURING STORAGE AT MICROELECTRONIC INDUSTRY ENTERPRISES^{1,3}Chernikov B.V., ²Kremer E.A., ³Borisova E.A.¹LLC «Gazprom VNNIGAZ», Moscow, bor-cher@yandex.ru;²Moscow Institute of Electronic Technology, Zelenograd, e-mail: kremerea@gmail.com;³Plekhanov Russian University of Economics, Moscow, e-mail: lb20062006@yandex.ru

Every day at the enterprises there are many events and processes that require reflection in various documents. During the manufacturing process, virtually every action needs to be recorded. A large volume of such actions leads to a constant increase in the number of created documents, and as a result, an increase in the volume of storages and databases allocated for these documents. All this leads to an increased load on the storage systems used at the enterprise. An increase in the number of documents entails an increase in the labor costs of the people who create these documents using traditional methods of generating documents when all text information is entered from the keyboard. Existing methods of document management and documentation do not have a significant impact on the current situation, and, therefore, are not sufficiently effective. This fact allows us to explain the use of lexicological synthesis, which will correct the current situation. Lexicological synthesis in the aspect of its application at microelectronics enterprises has a number of features. The paper discusses the methodology for reducing the volume of text documents, the creation of which makes it possible to reduce the already tangible results in the field of reducing the volume of text documents during storage.

Keywords: lexicological synthesis, technique, storage, slightly formalized, index sequence

Появление и распространение компьютеров, а также систем электронного документооборота не оказывает существенно влияния на ситуацию, сложившуюся на предприятиях в сфере документопроизводства. В настоящее время бумажные носители все еще являются основным способом хранения документов. Данная особенность предполагает, что на предприятии скапливаются архивы бумаг внушительных объемов.

Господствующее положение бумажных копий документов над электронными объясняется следующими причинами:

– порядок документооборота на предприятии, описанный в регламентирующих

эту сферу документах, таких как внутренние, государственные и ведомственные стандарты;

– меньшая защищенность цифровых копий документов из-за недостаточно быстрого распространения электронных средств защиты текстовых документов, например электронной подписи;

– инертность и консервативность регулирующих органов, партнеров и других организаций в работе с документами;

– высокая юридическая значимость.

Автоматизация документооборота достигается благодаря комплексному подходу к вопросу документооборота и процес-

са документирования на предприятии, где основой является документ [1]. Процесс создания, хранения, обработки и передачи документов является неотъемлемой и очень важной частью деятельности предприятия. Данный факт накладывает особые условия на этот процесс, поэтому к работе с информацией предъявляются требования:

- максимальное сокращение времени, необходимого для создания документа;
- минимально возможный объем создаваемого файла;
- максимальная формализация;
- автоматизация создания документов, независимо от степени их формализации.

Европейские стандарты, такие как MoReq или MoReq2, накладывающие дополнительные ограничения и условия на весь процесс создания документов, подразумевают соответствие требованиям, предъявляемым в нормативах [2, 3]. Повышение удобства работы с документами, в частности их формирования, также является требованием этих стандартов.

Слабоформализуемые документы являются основной частью документооборота на предприятиях микроэлектронной промышленности. Содержание данного типа документов сильно зависит от текущих условий и ситуации и необходимости учета всех особенностей формируемого документа. Слабоформализуемые документы – полнотекстовые, табличные либо смешанные документы, содержание которых существенно 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 [4].

Актуальность исследований в области создания, хранения, обработки и передачи документов на предприятиях обуславливается программой правительства Российской Федерации «Информационное общество» [5], а также постепенным переходом предприятий на электронный документооборот и растущим в связи с этим объемом электронных текстовых документов.

Целью данного исследования является создание модифицированной методики лексикологического синтеза, которая позволит сократить размер сохраняемых слабоформализуемых документов для использования на предприятиях микроэлектроники.

Особенности лексикологического синтеза документов

Лексикологический синтез – технология формирования текстовых фрагментов путем создания фраз на основе комплекса опорных слов, который формируется по результатам глубокого анализа текста совокупности документов путем связывания текстовых фрагментов с конкретным опорным словом, входящих в состав фраз или выражений формируемого текста [4]. Данный способ является одним из наиболее действенных способов для выполнения требований европейских и российских стандартов в области документооборота, который подразумевает автоматизацию создания документов.

В процессе подготовки к лексикологическому синтезу проводится унификация содержащейся в документе информации и создается набор фраз, которые могут находиться в документе. Для каждой такой фразы выбирается слово или словосочетание, наличие которого обозначает, что данная фраза присутствует в документе. В качестве такого слова или словосочетания выбирается то, которое несет основную смысловую нагрузку. Такие слова называются опорными. Совокупность опорных слов с учетом их взаимосвязей образует лексикологическую схему (лексонтологию) документа заданного типа. Лексикологическое дерево создается при помощи полного списка опорных слов, соответствующего заданной лексикологической схеме, и принятия во внимания вида информации, внедряемой в документ. Посредством последовательного выбора опорных слов происходит добавление необходимых формулировок в документ. В процессе формирования документа создается индексная последовательность, которая несет в себе результат выбора опорных слов. Сформированная индексная последовательность является готовым документом и содержит все нюансы содержания выбранного документа. Для возможности просмотра документа третьими лицами необходимо, чтобы у них было установлено необходимое программное обеспечение, в которое внедрено лексикологическое дерево, аналогичное тому, что использовалось во время формирования документа. Итоговый объем хранимого документа на практике оказывается в несколько раз меньше, чем документа, созданного с применением традиционных программных средств. Лексикологический синтез сводит возможность появления ошибок в тексте документа к минимуму, а также уменьшает время и трудозатраты, необходимые для формирования документа.

В качестве дополнительного преимущества обеспечивается легкость чтения готового документа посредством соответствующего программного обеспечения, тем самым не уступая традиционным способам формирования документов. Данные особенности являются важными факторами на предприятиях микроэлектроники.

Все сферы работы предприятия имеют конкретный набор документов. Комплект технической документации на предприятии включает в себя следующие группы [6]:

- конструкторские документы (Единая система конструкторской документации (ЕСКД) и нормативная документация (ГОСТы серии 2);

- технологические документы (Единая система технологической документации (ЕСТД), ГОСТы серии 3);

количество вариантов выбора является конечным. В процессе создания документа постоянная информация вносится автоматически. Унифицированная переменная информация вносится в процессе выбора нужных формулировок из общего множества опорных слов.

Методика сокращения объема документов

Информация, внедряемая в документ, меняется как по объему, так и по содержанию в зависимости от характера фиксируемых в документе событий. Наличие или отсутствие информации обуславливается различными факторами. Практически не существует условий, в которых возможно совпадение формируемых документов.

Рассмотрим методику сокращения объема документов при их создании на основе технологии лексикологического синтеза (рис. 1).

Рис. 1. Методика сокращения объема документов

- программные документы (на автоматизированные системы, регламентируются ГОСТами серии 34; на программные продукты – ГОСТами серии 19);

- эксплуатационные документы;

- ремонтные документы.

Информацию в любом документе можно разделить на переменную и постоянную информацию. Документный анализ позволяет выделить постоянную информацию, которая индивидуальна для каждого типа документов. Унификация текста позволит добавить переменную информацию, принадлежащую ограниченному множеству вариантов выбора. Такую информацию можно причислить к типу переменной унифицированной, потому что

1. Первый этап включает в себя глубокий информационный анализ документов [7].

Глубокий анализ структуры документов и информации, содержащейся в них, является основой для возможности применения лексикологического синтеза. Традиционное деление информации выделяет постоянную и переменную составляющие. К постоянной информации относят те данные, которые можно использовать без изменения в течение достаточно продолжительного периода. Переменная информация отражает фактические количественные и качественные характеристики какой-либо ситуации деятельности предприятия, которые необходимо закрепить в докумен-

те. Переменная информация традиционно включает в себя любые данные, которые необходимо вводить при каждом создании экземпляра документа заданного типа. Для применения лексикологического синтеза информация делится на следующие категории [8]:

- унифицированная постоянная информация (первый тип);
- унифицированная переменная информация (второй тип);
- переменная вводимая информация (третий тип);
- неунифицированная информация (четвертый тип).

Во время создания документа автоматически вводится в структуру первый тип информации. Эти данные хранятся в базе данных и готовятся заранее. Второй тип информации внедряется в документ посредством выбора требуемых опорных слов и соответствующих им формулировок, хранящихся в базе данных. Третий тип информации конкретизирует данные, соответствующие конкретному образцу документа. К четвертому типу относятся свободные формулировки, вводимые с клавиатуры, которые не предусмотрены опорными словами. Разделение информации на четыре типа по сравнению с классическим делением на два типа позволяет уменьшить объем информации, которая не предусмотрена заранее созданными формулировками, а также конкретизирует особенности события или факта на предприятии, описываемого в документе.

2. Следующий этап включает в себя построение лексонтологии и лексикологического дерева.

После анализа документов определяют формулировки, которым в соответствие ставятся опорные слова. Опорное слово должно иметь однозначное понимание, чтобы не возникло путаницы с выбором формулировок. Для подбора опорных слов применяют следующие критерии [4]:

- фонетический – опорное слово должно соответствовать звуковому строю языка документа;
- фоносемантический – звучание опорного слова должно вызывать ассоциации, непосредственно связанные с формируемым документом определенного вида;
- морфологический – составные, гибридные или сокращенные опорные слова, аббревиатуры или их идентификаторы должны как в полном объеме, так и составными его частями вызывать ассоциации, связанные с создаваемым документом;
- лексический и семантический – выбираемые или назначаемые опорные слова, их

совокупность или идентификатор должны базироваться на лексике документа;

- лексикографический – опорные слова, их совокупность или идентификатор должны легко читаться, смысл их должен быть понятен для любого составителя данного документа.

Количество опорных слов определяется каждым конкретным видом документа. Не должно быть переизбытка или недостатка опорных слов. Уменьшение количества опорных слов позволит уменьшить время формирования документа за счет укрупнения формулировок, соответствующих опорным словам, однако негативно скажется на изменчивости документа в целом в зависимости от ситуации. Это может быть оправдано, например для организационно-распорядительной документации, поскольку там используются устойчивые словосочетания, которые можно определить одним словом. Рост количества опорных слов позволит увеличить вариативность заполнения документа, но при этом может существенно возрасти время его формирования.

Исходя из всего этого, целесообразно выделить следующие критерии создания совокупности опорных слов [4]:

- лексико-смысловое единство;
- сбалансированная насыщенность множества опорных слов;
- смысловая ценность, обеспечивающая легкое запоминание смысловых цепочек опорных слов;
- функционально-стилевая принадлежность опорных слов к предметной области создаваемых документов, которая отражает специфику терминов, применяемых на предприятии.

После того, как множество опорных слов окончательно сформировано, на его основе строятся лексонтология и лексикологическое дерево 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 [4]. Основной задачей лексикологической схемы является отображение связи между опорными словами при формировании документа. Для лексонтологии используется древовидная структура данных, что позволяет наглядно показать взаимосвязи опорных слов с учетом вариативности создания документа. Учет связей между опорными словами и разбиение на 4 типа информации, содержащейся в документе, дает возможность построить лексикологическое дерево, пример которого приведен на рис. 2.

Рис. 2. Пример лексикологического дерева приказа об увольнении [4]

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

4. Четвертый этап – создание документа в автоматизированном режиме, когда пользователь выбирает необходимые опорные слова и вводит неунифицированную и переменную вводимую информацию.

5. Пятым этапом является создание индексной последовательности по алгоритму, который приведен на рис. 3. Индексная последовательность формируется последовательно, исходя из выбранных опорных слов, а в случае отсутствия варианта формулировки соответствует введенной с клавиатуры информации. После завершения создания основной части документа в индексную последовательность добавляются индексы подписей должностных лиц.

Модифицированный способ формирования индексной последовательности основан на использовании двоичных кодов. Выбор конкретного опорного слова отмечается «1», а пропуск опорного слова – «0».

На каждом уровне формируется своё двоичное число. Когда формирование числа на уровне завершено, оно переводится в десятичную систему. Это позволяет иметь для каждого уровня всего одно десятичное число. Номера уровней не указываются, поэтому уровни располагаются друг за другом. Уровни разделены пробелом. Если на уровне есть свободные формулировки, введенные с клавиатуры, то они располагаются после числа соответствующего уровня и заключаются в круглые скобки. По завершении формирования всех уровней свободные формулировки помещаются в одну строку друг за другом, сжимаются архиватором и добавляются в конец индексной последовательности. Все предварительные формулировки внутри последовательности удаляются и заменяются номерами, соответствующими положению в конкатенированной строке.

Рассмотрим пример формирования индексной последовательности на примере технологической инструкции. Лексонтология технологической инструкции представлена на рис. 4. Форма выбранного документа предусмотрена нормативным документом [9].

Рис. 3. Алгоритм формирования индексной последовательности [8]

Рис. 4. Лексонтология технологической инструкции

Итоговая индексная последовательность, созданная классическим методом:

1-3+«Плазма-12м-02» 2-1 3-1 4-1+4 5-1:1+2 6-1:1-2;2-4;4-6+2:3-5;5-7;6-8+3:140-260;140-170;150-170 7-1:1+5 8-1:1 9-1:1+2 A-1:3-5+2-1+3:5-10+4»

Итоговая индексная последовательность, созданная модифицированным методом:

4(«Плазма-12м-02») 1 1 9 1(3) 7(1-2 2-4 4-6;3-5 5-7 6-8;140-260 140-170 150-170) 1(17) 1(1) 1(3) 7(3-5;1;5-10;4)

Далее архивируется неунифицированная информация. В данном случае это строка выглядит как «Плазма-12м-02». Результат добавляется в конец последовательности, а сама последовательность, после исключения неунифицированного фрагмента, примет вид

4(«2») 1 1 9 1(3) 7(1-2 2-4 4-6;3-5 5-7 6-8;140-260 140-170 150-170) 1(17) 1(1) 1(3) 7(3-5;1;5-10;4)

Промежуточные результаты обоих методов можно сопоставить в таблице.

Сравнение классического и нового способов формирования последовательности

№ уровня	Классический способ	Новый способ
1	1-3+«Плазма-12м-02»	4(«Плазма-12м-02»)
2	2-1	1
3	3-1	1
4	4-1+4	9
5	5-1:1+2	1(3)
6	6-1:1-2;2-4;4-6+2:3-5;5-7;6-8+3:140-260;140-170;150-170	7(1-2 2-4 4-6;3-5 5-7 6-8;140-260 140-170 150-170)
7	7-1:1+5	1(17)
8	8-1:1	1(1)
9	9-1:1+2	1(3)
10	A-1:3-5+2-1+3:5-10+4	7(3-5;1;5-10;4)

Исходя из приведенных выше результатов можно сделать вывод, что новый метод формирования индексной последовательности выигрывает у классического в плане объема формируемого документа.

6. На следующем этапе в индексную последовательность внедряются индексы графических компонентов документа, к которым можно отнести логотипы, схемы и т.п.

7. Седьмой этап включает в себя сжатие последовательности с помощью архиватора.

Выводы

В результате исследования получены следующие результаты.

1. Разработана методика на основе лексикологического синтеза для сокращения объема документов, хранимых в электронном виде, и приведен пример ее применения для создания технологической инструкции, что доказало эффективность данной методики.

2. Разработанная методика описывает последовательность этапов, в которых описаны действия, выполняемые для создания документа. Методику можно применять на предприятиях микроэлектроники.

3. Исследование выявило следующие проблемы, которые будут решены в дальнейших исследованиях:

– при наличии большого объема информации, введенной с клавиатуры, объем ин-

дексной последовательности может существенно увеличиться;

– отсутствие учета неиндексированных графических материалов, которые могут быть добавлены в процессе создания документа.

Список литературы

1. Федеральный закон от 27 июля 2006 года № 149-ФЗ «Об информации, информационных технологиях и о защите информации» // Собрание законодательства Российской Федерации. 31.07.2006. № 31 (ч. 1). ст. 3448.
2. Model Requirements for the Management of Electronic Records (MoReq 2). France: European Communities. 2008. 212 p.
3. Modular Requirements for Records Systems (MoReq 2010). CEEA-CEE-CEE, Bruxelles-Luxembourg. 2011. P. 520.
4. Черников Б.В. Лексикологический синтез документов в комплексах информационных систем. М.: ИД «ФОРУМ», 2017. 336 с.
5. 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 (ч. 2). ст. 2159.
6. Глаголев В.А. Проблемы стандартизации и повышения качества разработки технической документации на современную промышленную продукцию приборостроения // Менеджмент качества. 2008. № 4. Ч. 1. С. 68–78. 2009. № 1. Ч. 2. С. 57–65.
7. Черников Б.В., Кремер Е.А. Анализ информации текстовых документов предприятий микроэлектронной промышленности // Современные наукоемкие технологии. 2018. № 5. С. 168–172.
8. Черников Б.В., Кремер Е.А. Сокращение объема индексной последовательности при лексикологическом синтезе слабоформализуемых документов // Информатизация и связь. 2018. № 6. С. 58–64.
9. ГОСТ 3.1105-2011. Единая система технологической документации (ЕСТД). Формы и правила оформления документов общего назначения. М.: Стандартинформ, 2012. 24 с.