УДК [005.2:004] (076.5)

Б.В. Черников
Национальный исследовательский университет «Высшая школа экономики»
Профессор кафедры «Архитектура программных систем»

ЛЕКСИКОЛОГИЧЕСКИЙ СИНТЕЗ – ПРОГРЕССИВНАЯ ТЕХНОЛОГИЯ СОЗДАНИЯ ТЕКСТОВЫХ ДОКУМЕНТОВ

Излагается сущность лексикологического синтеза слабоформализуемых текстовых документов, обеспечивающая существенное сокращение трудозатрат на создание документов, повышение их качества, снижение объемов хранения и рост защищенности при передаче по каналам связи.

Ключевые слова: лексикологический синтез, слабоформализуемый документ, опорное слово.

Современная деятельность организации любого профиля неразрывно связана с реализацией управленческих решений, процесс выработки которых опирается на информационное обеспечение. В связи с этим чрезвычайно важен обмен своевременной и точной информацией, который чаще всего выполняется с помощью предварительно подготовленных документов, сопровождающих производственные процессы.

Функции управления организацией выражаются посредством соответствующих документов. В условиях неуклонного усложнения деятельности особое внимание обращает на себя возрастание требований полноты и своевременности, предъявляемых к управленческой информации. Офисные функции, определяющие подготовку необходимых документов для выработки управленческого решения, реализуются в организациях в рамках документационного обеспечения управления, включающего в себя процессы документирования и организации работы с документами, причем если задачи второго процесса автоматизированы специализированными системами, то вопросы автоматизации создания, оформления и изготовления документов развиты недостаточно и требуют особого рассмотрения.

Характерно, что при автоматизации процесса подготовки документов руководители, как правило, не чувствуют преимуществ внедрения новых технологий. Особую значимость этот вопрос приобретает в случаях создания документов непосредственно их исполнителями, что чаще всего имеет место. Пока же в организациях и на предприятиях многие документы, в основном, создаются на основе прямого ввода текста с клавиатуры с использованием возможностей текстовых процессоров, что обусловливает возможность появления ошибок при подготовке документов исполнителями. Передача документов по каналам связи и их хранение сопряжены со значительностью объема и недостаточной защищенностью.

На современном этапе процесс документирования информации, поддерживающей производственные процессы организаций, должен отвечать следующим требованиям:

- данные должны быть максимально формализованы в целях обеспечения автоматизированной обработки сведений, содержащихся в документе;
- создание документов должно занимать минимум времени при сохранении требований к информации, необходимых для поддержки процесса принятия управленческого решения;
- в условиях слабой формализации, характерной для текстовых документов, используемых в системах поддержки основной деятельности и процессов принятия решений, необходимо предварительное приведение содержания документа к виду, пригодному для автоматизированного формирования конкретного экземпляра документа.

Данные требования соответствуют как отечественным рекомендующим и нормативным документам, так и положениям европейского стандарта MoReq2. Кроме указанных тре-

бований, MoReq2 большое внимание уделяет простоте пользования и производительности, что гармонирует с требованиями, определяемыми стандартом открытых систем.

На основе результатов проведенного анализа информационного состава документов выявлено, что при традиционной классификации информации, когда рассматриваются всего два ее вида – постоянная и переменная, в документах преобладает переменная информация, в то время как постоянная информация содержится в достаточно небольших объемах (табл. 1).

Таблица 1

Информационный состав документов при традиционной классификации информации

Группа документов	Постоянная информация, %	Переменная информация, %
Организационно-распорядительные документы	13 – 18	82 – 87
Документы промышленного предприятия	3 – 16	84 – 97
Документы высшего учебного заведения	3 – 18	82 – 97
Документы лечебного учреждения	6 – 18	82 – 94

Анализ состава информации, содержащейся в документах организаций, показывает, что доля постоянной информации, которая может быть заблаговременно внесена в трафареты и шаблоны формируемых документов, не является достаточно высокой для получения значимого эффекта при существенной вариации содержания документов. В среднем по рассматриваемым группам документов доля постоянной информации составляет:

- для организационно-распорядительных документов – 14,8%;
- для документов промышленного предприятия – 8,1%;
- для документов высшего учебного заведения – 8,4%;
- для документов лечебного учреждения – 8,7%;

Обращают на себя внимание достаточно близкие значения объемов постоянной информации для организационно-распорядительных документов предприятия и существенные различия этого показателя – в документах производственно-технологического назначения.

Анализ исходной трудоемкости формирования документов организаций и состава содержащейся в них информации показывает, что относительно невысокие трудозатраты на создание организационно-распорядительных документов могут объясняться более высоким уровнем типизации текста [1, 2]. В то же время создание документации, относящейся к другим группам и обладающей более низким уровнем типизации (о чем свидетельствуют малые значения объемов постоянной информации) требует более значительных трудозатрат персонала [3, 4].

Выход из создавшегося положения может быть найден только в пересмотре традиционных процедур подготовки документов. Значительное повышение эффективности процесса создания документов возможно благодаря использованию лексикологического синтеза, который апробирован на создании документов различного назначения в организациях широкого спектра деятельности. Апробация показала, что лексикологический синтез весьма эффективен при создании слабоформализуемых документов, причем уровень выигрыша пропорционален частоте повторяемости документа: чем чаще формируется документ, тем больше выигрыш.

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. При анализе состава значительное число документов, сопровождающих производственные процессы, можно отнести к категории слабоформализуемых.

Принцип лексикологического синтеза предполагает проведение предварительного анализа совокупности документов определенного вида, на основе которого выделяется устойчивый набор формулировок. Каждой формулировке документа ставится в соответствие опорное слово, выбор которого однозначно определяет наличие конкретной формулировки в документе. Если выделенный фрагмент текста документа содержит значительное количество строк и всегда присутствует в документе в строго определенной последовательности построения предложений, то такой фрагмент текста может определяться одним опорным словом.

Традиционный подход, который применяется в настоящее время при использовании типизации и трафаретизации документов, предлагает деление информации на постоянную и переменную. Однако при разработке процесса автоматизированного формирования документов на основе лексикологического синтеза целесообразно использовать более глубокую классификацию информации, а именно принимать во внимание не две категории (постоянная и переменная информация), а четыре (рис. 1). В этом случае целесообразно классифицировать информационные потоки документов по следующим категориям:

- унифицированная постоянная информация, подготовленная заранее и хранимая в базе данных или содержащаяся в тексте программы, которая автоматически внедряется в формируемый документ программными средствами. К этому типу относится постоянная информация (например, наименование документа) и редко меняющаяся (наименования структурных подразделений, список персонала, перечень разделов документа и т.п.);
- унифицированная переменная информация, содержащая стандартизированные и формализованные данные, хранимая в базе данных и вводимая при формировании документа путем выбора требуемых формулировок. Этот тип информации включает в себя именно те формулировки, которые предлагаются исполнителю для выбора при формировании документа;
- переменная вводимая информация, подчиненная определенным требованиям по способу представления данных и содержащая конкретизирующие сведения, как правило, для конкретного экземпляра документа (например, табличные данные, отдельные фамилии, характеристики оборудования, данные по рекомендуемым режимам работы, оценки при проведении контрольных мероприятий и т.п.) и вводимая с клавиатуры непосредственно при подготовке документа;
- неунифицированная информация, содержащая свободные формулировки и вводимая при необходимости прямым набором текста с клавиатуры.

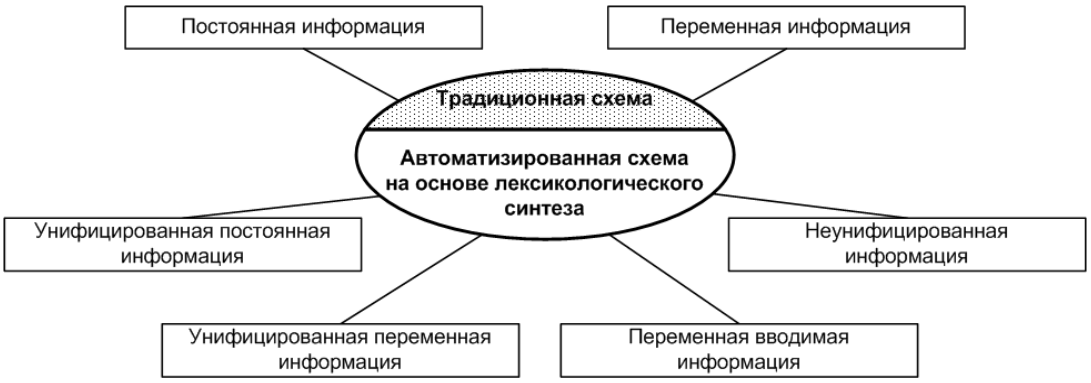

Рис. 1. Категории информации в схемах формирования документов

Учитывая значимость роли процесса документирования в совершенствовании документационного обеспечения управления деятельностью организаций, разработана методология информатизации документационного обеспечения и подготовки слабоформализуемых документов, циркулирующих в системе управления и непосредственно относящихся к проведению основной деятельности организаций и предприятий, а также прикладных аспектов ее реализации.

Слабоформализуемые документы, обеспечивающие принятие решений по наблюдаемым событиям или фактам, формируются с помощью автоматизированного лексикологического синтеза путем обхода лексикологического дерева [5, 6, 7].

Каждой формулировке документа ставится в соответствие основное слово, выбор которого однозначно определяет наличие конкретной формулировки в документе. Такие слова являются опорными, из них составляется лексикологическая схема формируемого документа. Таким образом, опорными называются слова, выбор которых при создании документа однозначно определяет наличие в документе конкретных формулировок, связанных с этими словами.

Множество взаимно зависимых опорных слов в совокупности определяет последовательность обхода маршрута формирования документа. На основе предварительного анализа структуры документа выявляются основные разделы, которые должны или могут присутствовать в документе. Условные наименования таких разделов составляют основу синтезируемой совокупности опорных слов. В рамках каждого зафиксированного раздела документа выявляются составные элементы, которые должны или могут входить в состав раздела (слово, фраза, текстовой фрагмент). Для каждого подобного составного элемента определяется опорное слово (или их совокупность), выбор которого в последующем однозначно будет определять внедрение в документ соответствующего фрагмента. Если фрагмент текста документа содержит значительное количество строк и всегда присутствует в документе в строго определенной последовательности построения предложений, то данный фрагмент текста может определяться одним опорным словом. Однако в случаях, когда текст документа формируется из предложений, не фиксированных в строго определенной последовательности, и в каждом заново создаваемом документе наблюдаются вариации построения текста, опорных слов будет столько, сколько необходимо для однозначного определения каждого конкретного предложения или словосочетания.

Полный перечень опорных слов с учетом их взаимосвязей образует лексикологическое дерево документа, «прохождение» по ветвям которого обеспечит выбор формулировок, используемых в документе. При этом выбор тех или иных опорных слов будет означать необходимость внедрения в документ совершенно конкретных вариантов текстовых фрагментов. Фактически текст документа формируется путем выбора необходимых заготовок из числа сохраненных формулировок. Структура лексикологического дерева сходна с композицией текста документа. Степень ветвления лексикологического дерева зависит от объема множества вариаций текста документа, определяемых его сложностью и различием документируемых ситуаций.

В отличие от «прямого» лексикологического дерева, в котором в обязательном порядке последовательно выбираются все его компоненты, более общим случаем является вариант выбора опорных слов с отсечением, когда выбор очередного опорного слова зависит от того, какое опорное слово было выбрано на предыдущем шаге цикла. При этом для каждого конкретного экземпляра документа определенного вида формируется маршрут выбора опорных слов, отсекающий ряд боковых ветвей (рис. 2, где утолщенной линией указан маршрут выбора совокупности опорных слов, входящих в состав лексикологического дерева и выбираемых в процессе обхода дерева при формировании документа). В результате сформированное по такому принципу лексикологическое дерево обеспечивает поддержку не одного экземпляра, а совокупности документов определенного вида.

В качестве опорного слова могут выступать различные части речи, определяющие сущность предписываемого действия.

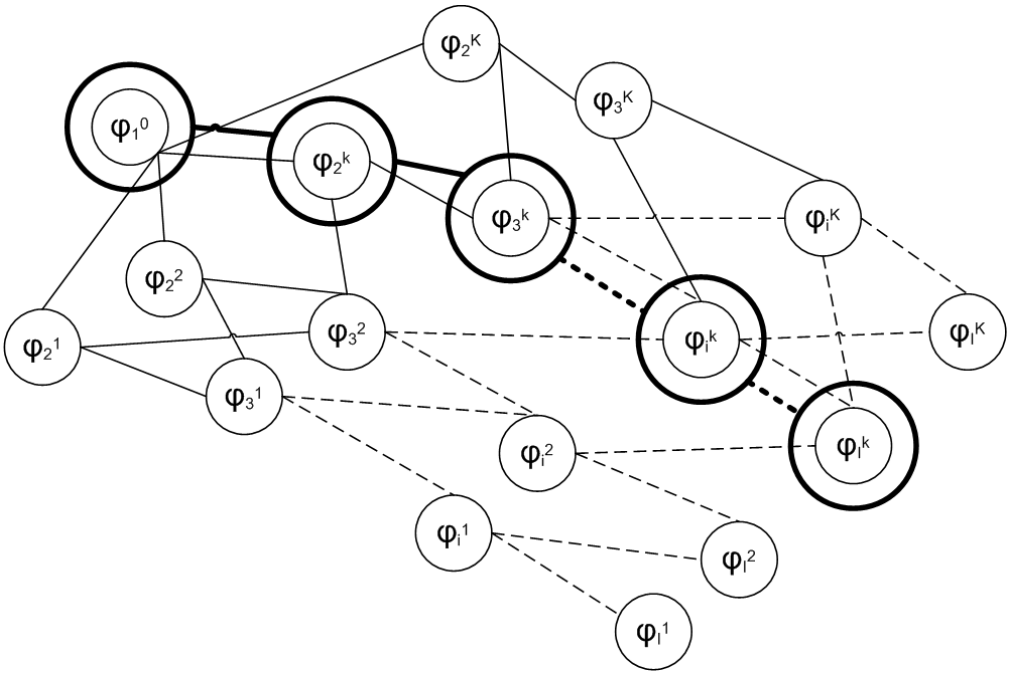

Рис. 2. Модель формирования документа при использовании лексикологического дерева с отсечением

Модель формирования документа при использовании дерева с отсечением при ветвлении подобного типа можно представить следующим образом (логическим суммированием, характеризующим образование конкатенаций текстовых фрагментов документа, учтено, что выбираются не все опорные слова, а лишь некоторые из них, хотя все они, безусловно, принадлежат множеству опорных слов документа данного вида):

$$D^{\mathrm{B}} \Rightarrow \sum_{i=1}^{I^{\mathrm{B}}} (\varphi_i \,|\, \varphi_{i-1}) \text{ при } \varphi_i \in \boldsymbol{\Psi}^{\mathrm{B}},$$

где φ_i – текущее опорное слово; I^{B} – количество опорных слов для документа D^{B} конкретного вида; i – условный номер (индекс) текущего опорного слова; $\boldsymbol{\Psi}^{\mathrm{B}}$ – множество опорных слов документа данного вида.

При генерации лексикологической схемы, представляющей собой своеобразную онтологическую модель документа, и лексикологического дерева следует соблюдать принцип управления лексическими конструкциями и учитывать онтологическую относительность статуса опорных слов. Представления об онтологической относительности высказывались в концепции языковых каркасов Карнапа [8], развивающей идею многоступенчатого исчисления предикатов. Опорное слово должно быть уникальным для конкретной конструкции, а при необходимости – уточняться другими опорными словами, иначе выбор требуемого текстового фрагмента может быть определен неверно. Уточнение одного опорного слова другим образует их иерархическую подчиненность в структуре лексикологического дерева документа определенного вида.

Формирование лексикологической схемы и лексикологического дерева проводится на основе анализа связи опорных слов, образуя маршрут (траекторию) их выбора при формировании документа. Лексикологическая схема позволяет установить взаимосвязи опорных слов, учитывая, что в различных случаях может наблюдаться существенное изменение маршрута формирования документов, что и определяет возможную вариативность отдельно взятых экземпляров. Пример лексикологической схемы приведен на рис. 3 для протокола эндоскопического осмотра при проведении гастроскопии.

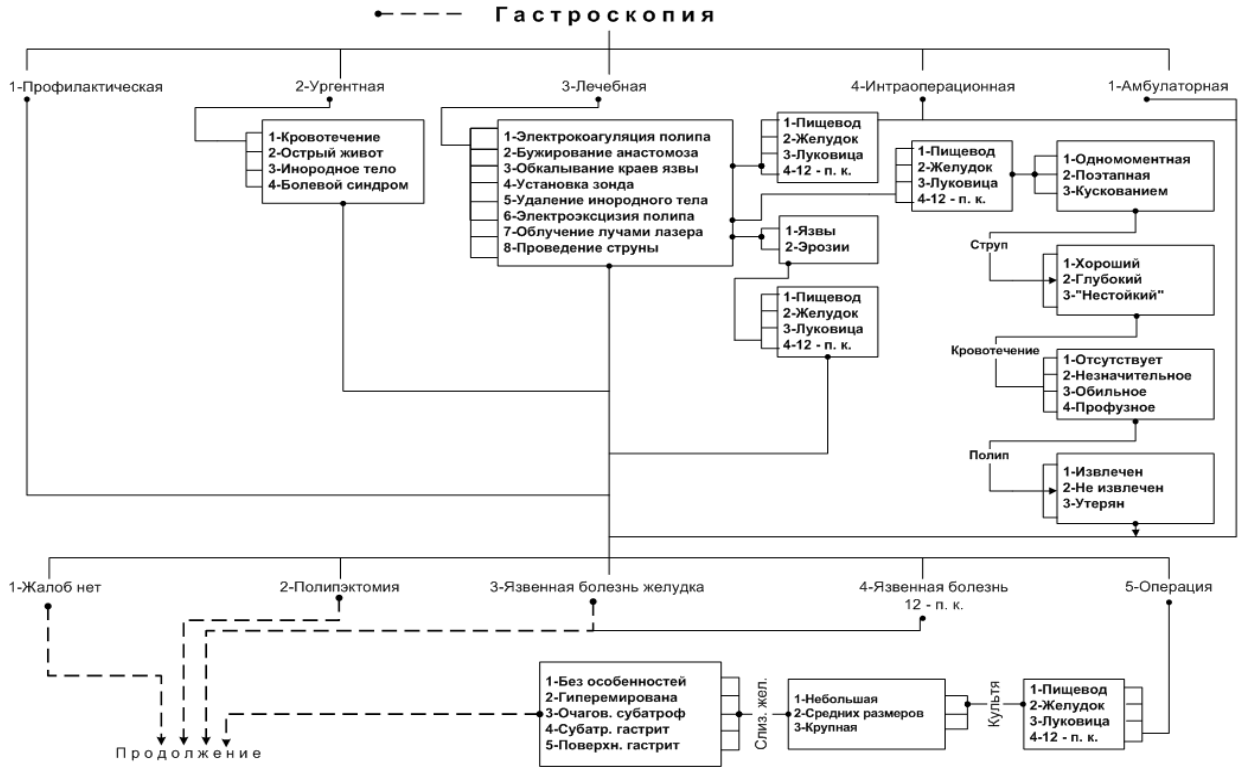

Рис. 3. Пример лексикологической схемы протокола гастроскопии

На этапе автоматизированного формирования документов (период эксплуатации систем) используется база данных комплекса формируемых документов, содержащая реквизиты, формы, лексикологические деревья и комплекты опорных слов документов. При формировании конкретного экземпляра документа после выбора первого опорного слова в документ внедряется фрагмент, соответствующий этому опорному слову. Процесс выбора опорных слов повторяется по всему маршруту формирования документа с учетом взаимосвязей опорных слов. По завершении процесса выбора опорных слов в унифицированную форму внедряются все положенные реквизиты и оформляется конкретный экземпляр документа.

Полный перечень опорных слов с учетом их взаимосвязей образует лексикологическое дерево документа, «прохождение» по ветвям которого обеспечит выбор формулировок, используемых в документе.

В процессе исследования сформирована модель разработки автоматизированной технологии создания слабоформализуемых текстовых документов на основе лексикологического синтеза (рис. 4). В соответствии с решаемыми на каждом этапе задачами в структуре модели предусмотрены четыре стадии разработки: аналитическая, унификационная, информационная и процедурная. На аналитической стадии выделяется комплекс документов организации, для которых будет разрабатываться технология автоматизированного формирования документов. Процесс приведения документов к единообразию по форме и содержанию реализуется на унификационной стадии, значение которой велико, поскольку унификация обеспечивает возможность совместного и многократного использования конструкций различных документов и интероперабельность, т.е. независимость от технической и программной платформы их создания и обработки, что создает предпосылки к экономии времени и материальных затрат. Информационная стадия разработки технологии лексикологического синтеза слабоформализуемых документов предназначена для построения своеобразной инфологической модели данных. На процедурной стадии проводится подготовка программных компонентов, предназначенных для непосредственной реализации процесса создания документов и интеграции системы формирования с информационной системой предприятия.

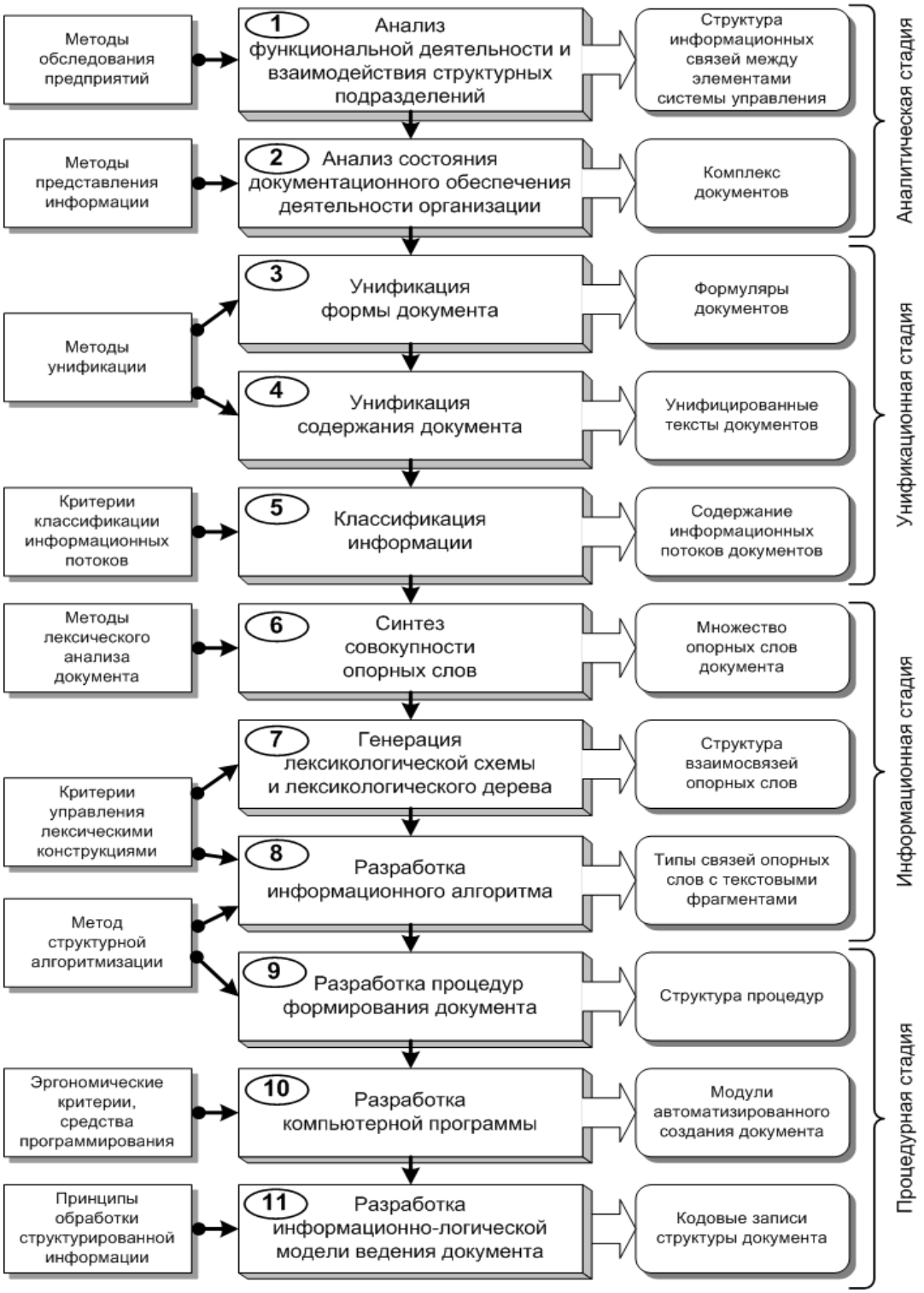

Рис. 4. Модель разработки технологии создания слабоформализуемых документов

Лексикологический синтез слабоформализуемых документов позволяет существенно сократить нагрузку на системы хранения информации. При составлении документа формируется индексная последовательность, подвергаемая дополнительному сжатию для создания сохраняемого информационного пакета, а при воссоздании документа производится расшифровка этого пакета с использованием организованного цикла восстановления фрагментов на основе согласованной базы данных лексикологического дерева [9]. Анализ эффективности предлагаемого метода обработки слабоформализуемых документов при организации их хранения по-

казывает возможность значительного (в десятки и даже сотни раз) сокращения объемов хранимой информации по сравнению с прямым сохранением или сжатием информации.

Графическое отображение показателей объема сохраняемых документов приведено для организационно-распорядительных документов и документов промышленного предприятия – на рис. 5, для документов высшего учебного заведения и документов лечебного учреждения – на рис. 6.

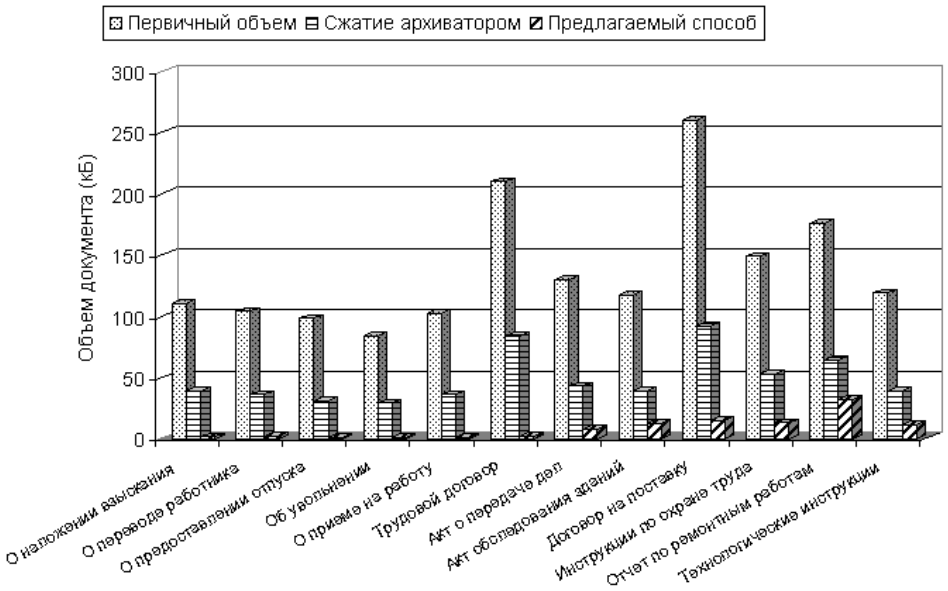

Рис. 5. Объем сохраняемых организационно-распорядительных документов и документов промышленного предприятия

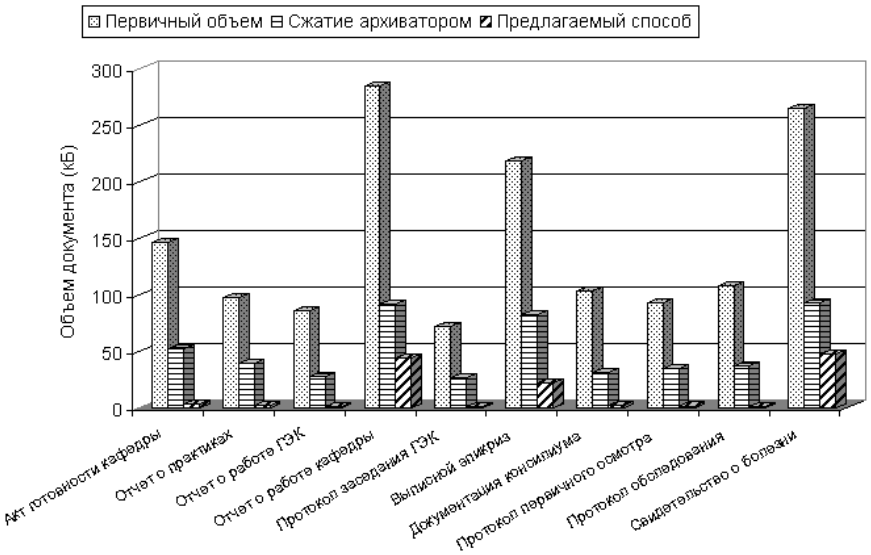

Рис. 6. Объем сохраняемых документов высшего учебного заведения и лечебного учреждения

Внедрение автоматизированного способа лексикологического синтеза в практику формирования слабоформализуемых текстовых документов обеспечивает смещение групповых центров аккумуляции документарных компонентов из областей неунифицированной информации в сторону унифицированных элементов, что предопределяет возможность автоматизированного выбора унифицированных формулировок в автоматизированной схеме (рис. 7).

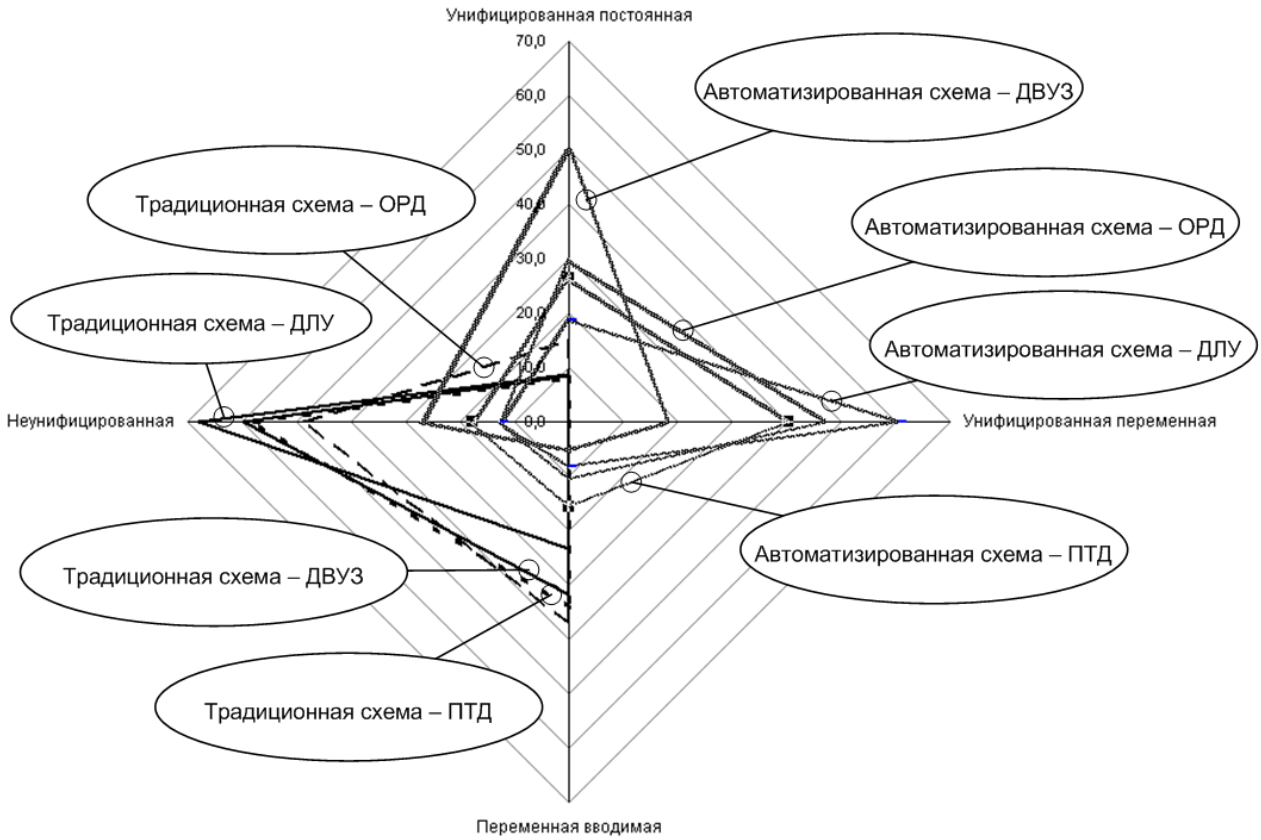

Рис. 7. Результаты реструктуризации информационных потоков документов по подсистемам документов

Обращает на себя внимание значительность роста унифицированной информации по сравнению с традиционной схемой формирования документов (в 5 раз в организационно-распорядительных документах, в 8 раз в производственно-технологической документации и документации высших учебных заведений, в 9 раз в документации лечебных учреждений) и значительное снижение объемов переменной вводимой и неунифицированной информации (табл. 2).

Таблица 2

Усредненные показатели изменения доли информационных компонентов документов

Показатель	ОРД	ПТД	ДВУЗ	ДЛУ
Рост удельного веса унифицированной информации	5,2	8,2	8,1	9,1
Сокращение объемов переменной вводимой информации	3,5	2,2	6,2	2,9
Снижение объемов неунифицированной информации	3,9	3,2	2,2	5,4

Анализ трудоемкости при изменении условий формирования документов демонстрирует в среднем более чем двукратное снижение трудозатрат при использовании предлагаемых технологий в организациях различного профиля деятельности.

Графически динамика трудоемкости создания организационно-распорядительных документов и документов промышленного предприятия при внедрении лексикологического синтеза отображена на рис. 8.

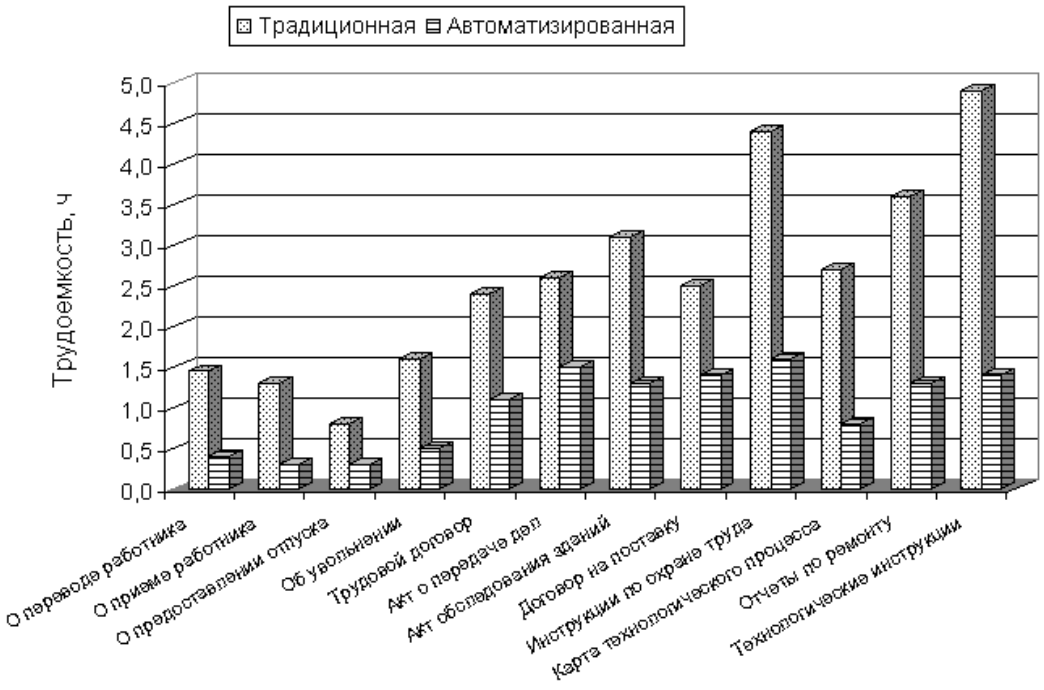

Рис. 8. Трудоемкость создания организационно-распорядительных документов и документов промышленного предприятия

В целом разработка и исследование возможностей лексикологического синтеза текстовых документов показали высокую эффективность и перспективность этой технологии создания документов. Дополнительно следует заметить, что помимо сокращения трудозатрат, необходимых для формирования слабоформализуемых документов в организациях, одновременно повышается качество формируемых документов.

Библиографические ссылки

1. Черников Б.В. Методология автоматизации документационного обеспечения управления / Б.В. Черников. – М.: ГУУ, 2002. – 12 с.

2. Черников Б.В. Принцип лексикологического синтеза в технологии создания текстовых документов / Б.В. Черников // Секретарское дело. – 2000 – № 1. – С. 47-49.

3. Черников Б.В. Технология автоматизированного формирования слабоформализуемых документов / Б.В. Черников, А.М. Карминский // Сб. тр. Третьей межд. конф. «Управление развитием крупномасштабных систем» MLSD’2009. – М.: Ин-т проблем управления им. В.А. Трапезникова РАН. – С. 398-408.

4. Черников Б.В. Формирование электронного документа / Б.В. Черников // Служба кадров. – 2007. – № 11. – С. 97-102.

5. Черников Б.В. Лексикологический синтез слабоформализуемых документов / Б.В. Черников // Информационные технологии и вычислительные системы. – 2009. – № 4. – С. 104-115.

6. Черников Б.В. Способ автоматизированного лексикологического синтеза документов / Б.В. Черников. – Патент РФ № 2253893, 2005.

7. Черников Б.В. Технологии подготовки документов на основе кибернетических методов / Б.В. Черников. – М.: Финансы и статистика, 2009. – 208 с.

8. Карнап Р. Значение и необходимость. Исследования по семантике и модальной логике / Р. Карнап. – М.: ЛКИ, 2007. – 384 с.

9. Черников Б.В. Способ преобразования слабоформализуемых документов для минимизации их объема при хранении / Б.В. Черников. – Патент РФ № 2413985, 2011.