ISSN 1992-2264
Том 3 Выпуск 4 Октябрь–декабрь 2009

Российская академия наук

# ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ

**ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ**

**Научный журнал Отделения нанотехнологий и информационных технологий Российской академии наук**

Издается с 2007 года
Журнал выходит ежеквартально

**Учредители:**
**Российская академия наук**
**Институт проблем информатики Российской академии наук**





**Адрес редакции:**

Москва 119333, ул Вавилова 44, корп 2, ИПИ РАН,
редакция журнала «Информатика и её применения»
Тел 8(499)135-86-92, e-mail rust@ipiran.ru

**Журнал «Информатика и её применения» включен в «Перечень ведущих рецензируемых научных журналов и изданий, в которых должны быть опубликованы основные научные результаты диссертации на соискание ученой степени доктора и кандидата наук», утвержденный ВАК**

Журнал зарегистрирован в Федеральной службе по надзору в сфере связи и массовых коммуникаций 31 марта 2009 г
Свидетельство о регистрации средства массовой информации ПИ № ФС77-35837

Подписной индекс журнала в каталоге «Пресса России» 88018 (годовая подписка)

# Информатика и её применения

Том 3 Выпуск 4 Год 2009


## СОДЕРЖАНИЕ




Выпускающий редактор *Т. Торжкова*, Технический редактор *Л. Кокушкина*
Художественный редактор *М. Седакова*
Сдано в набор 26.10.09. Подписано в печать 30.11.09. Формат 60 x 84 / 8
Бумага офсетная. Печать офсетная. Усл.-печ. л. 14,0. Уч.-изд. л. 10,5. Тираж 200 экз.
Заказ № 2107
Издательство «ТОРУС ПРЕСС», Москва 121614, ул. Крылатская, 29-1-43
torus@torus-press.ru, http://www.torus-press.ru
Отпечатано в ППП «Типография «Наука» с готовых диапозитивов, Москва 121099, Шубинский пер., д. 6.




# ТЕХНОЛОГИЯ ХРАНЕНИЯ СЛАБОФОРМАЛИЗУЕМЫХ ДОКУМЕНТОВ НА ОСНОВЕ ЛЕКСИКОЛОГИЧЕСКОГО СИНТЕЗ


Б. В. Черников[1]



**Аннотация:** Рассмотрена технология хранения слабоформализуемых документов, создаваемых с помощью лексикологического синтеза. Технология предусматривает формирование сохраняемых индексных последовательностей, содержащих индексы форм документов и их содержательных компонентов. Благодаря одновременной подготовке документов и созданию сохраняемых индексных последовательностей дополнительно обеспечивается экономия времени. Эксперименты показали эффективность подхода для документов, создаваемых в интересах управления различными видами деятельности.



**Ключевые слова:** слабоформализуемый документ; лексикологический синтез; индекс; индексная последовательность; сжатие


## 1 Введение

На протяжении жизненного цикла информационных систем, использующих персональные компьютеры как средства создания документов, возникают проблемы, связанные с хранением информации, поскольку при постоянно возрастающем числе документов требуются большие объемы памяти. Объем корпоративной информации увеличивается ежегодно, причем рост объемов сохраняемой информации достаточно значителен. Согласно исследованиям компании IDC, рост объемов хранимых и обрабатываемых данных может достигать более 70% в год [1]. Росту объемов сохраняемых данных способствуют как требования различных нормативных актов, так и внутрикорпоративные стандарты, устанавливающие необходимость длительного хранения некоторых видов информации.

Очень часто обработка, хранение и архивирование данных рассматриваются с сугубо технической или экономической точки зрения, причем правовые риски этих процессов явно недооцениваются. Важность сохранения документов в первую очередь определяется значимостью самих документов.

Государство, правительственные органы и деловые круги рассматривают возможность сохранения различного рода документов, корреспонденции, записей осуществления транзакций, счетов, контрактов и другой важной информации в качестве реализации прав на доказательство [2].

## 2 Технологии преобразования документов при хранении

В целях сокращения объемов сохраняемой и формации при организации хранения данные мог подвергаться сжатию, для чего используются апп ратные и программные методы. Базовым принц пом, лежащим в основе методов сжатия данны является устранение избыточности, содержащей в сохраняемой информации.

Производители аппаратных средств сжатия да ных используют, как правило, стандартизованны алгоритмы (например, Лемпела–Зива LZ1), позво ляющие ускорить процесс сжатия [3, 4]. Функци сжатия информации широко используется в сре ствах резервного копирования данных. Преобраз вание в этих случаях обычно осуществляется непо средственно в устройствах хранения (например, ленточных библиотеках), что снижает нагрузку н серверное оборудование. Кроме того, в [5] при водятся сведения о результатах двойного сжати применяемого на основе различающихся алгорит мов в системах компаний Avamar (*Axion*) и Rockso (*Blocklets*). В таких системах устранение избыточности осуществляется на байтовом уровне: сначал последовательно удаляются повторяющиеся цепоч ки, а затем агрегируются полученные данные.

Программные методы сжатия достаточно попу лярны и чаще всего используются в виде програм архиваторов [6–9], позволяющих снизить разме сохраняемых файлов благодаря специальным ме тодам обработки. Среди программных методо


[1]ООО «АНТ-Информ», Москва, bor-cher@yandex.ru

сжатия различаются методы сжатия без потерь (применяются для сокращения объемов хранения текстовых документов) и методы сжатия с потерями, позволяющие снизить размеры графических, аудио- и видеофайлов за счет исключения при обработке несущественной информации. Сжатие данных при архивировании позволяет сократить размер файла в несколько раз, однако степень сжатия для файлов разных типов различна. Так, файлы формата PDF содержат сжатую информацию уже в первичном виде и потому практически не сжимаются архиваторами дополнительно.

Вопросы разработки алгоритмов сжатия, которые лежат в основе программных методов, исследуются специалистами в поисках путей совершенствования способов сокращения объемов хранимой информации [10–17].

Работа [17] посвящена изложению особенностей технологии *Venom*, позволяющей сжать строки данных для снижения потребности в дисковом пространстве и тем самым повысить эффективность ввода/вывода в базе данных DB2. Технология построена на сканировании таблиц баз данных, поиске повторяющейся информации и построении словарей, которые присваивают повторяющимся объектам короткие числовые ключи.

В работе Богатова [16] рассмотрена возможность увеличения степени сжатия на основе методов стохастического контекстного моделирования с применением модели разреженных контекстов.

Кадач в [11] предложил способ сжатия на основе кодирования слов, содержащихся в тексте или гипертексте, значениями их позиций в сохраняемом словаре. К недостатку этого метода следует отнести необходимость обязательного хранения всех модификаций слов, вызванных их видоизменением при связывании (склонением или спряжением), а также чрезмерную избыточность словарей.

В работе [11] предложен также способ повышения степени сжатия на 5%–10% благодаря адаптивному изменению алфавитного порядка в алгоритме обобщенных интервальных преобразований. Другая модификация интервального кодирования, когда каждая буква исходного слова заменяется числом букв с большими номерами, разделяющих текущее и предыдущее включение буквы, была предложена Арнавутом и Магливерасом [12].

Работа [14] посвящена рассмотрению способов повышения стандартных средств сжатия данных на основе преобразования индексов слов, образующих информационную посылку. В обзоре алгоритмических методов моделирования в целях сокращения объемов текстов [10] авторами рассмотрен достаточно широкий спектр подходов к сжатию информации, среди которых наиболее близко к технологии, рассматриваемой в основной части данной статьи, находятся словарные кодировщики. Махони в работе [13] рассматривает способ повышения скорости сжатия информации на основе применения нейросетей.

Наиболее привычным способом перевода электронных документов в хранимую версию является их сохранение на выбранных носителях информации непосредственно из программных сред, в которых осуществлялось создание данных документов. К недостаткам такого способа следует отнести необходимость полнотекстового сохранения документа со всеми его неотъемлемыми компонентами и атрибутами, включая служебную информацию, присущую программным средам. Например, для документов, создаваемых в Microsoft Word, такая информация имеет достаточно большой объем («пустой» документ, содержащий лишь служебную информацию текстового процессора, при сохранении на диске занимает более 35 кБ). Большинство деловых документов создаются с использованием бланков, содержащих графические компоненты как неотъемлемую часть документа. Наличие в документе графических компонентов (например, изображение на бланке цветного логотипа организации) может приводить к увеличению общего объема документа более чем на сотни килобайт.

Архивирование файлов, позволяющее сократить их объем, широко применяется на практике, однако для создания файла-архива требуется проведение дополнительной операции сжатия (после сохранения на носителе информации) только по завершении процесса создания документа. Итак, все рассмотренные выше алгоритмические методы сжатия информации в обязательном порядке требуют наличия *уже готового документа*, который *после его создания* должен подвергаться дополнительной обработке в целях сокращения объема, подлежащего сохранению.

## 3 Цель исследования

Целью настоящего исследования являлась разработка новой технологии преобразования не только информации, содержащейся в документах, но и документов в целом, обеспечивающей сокращение объемов хранимых данных.

Для сокращения времени обработки документа и исключения дополнительных операций целесообразно совместить процессы автоматизированного формирования слабоформализуемых документов на основе лексикологического синтеза и их





преобразования для последующего хранения в сокращенном объеме и автоматизированного восстановления.

Разработанная автором технология апробирована на подсистеме организационно-распорядительных документов, отчетных документов кафедрального уровня высшего учебного заведения, а также ряда учетных и отчетных документов медицинской направленности.

## 4 Преобразование слабоформализуемых документов при хранении

Сокращение объема информации, содержащейся в создаваемых документах и сохраняемой на носителях, возможно благодаря использованию особенностей способа лексикологического синтеза, применение которого весьма эффективно при создании слабоформализуемых документов.

*Слабоформализуемые документы* — полнотекстовые, табличные или смешанные документы, содержание которых существенным образом связано с произвольной, меняющейся в каждой 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.

Использование лексикологического синтеза для создания слабоформализуемых документов сопряжено с глубоким предварительным анализом документов определенного вида. На первый взгляд подобную задачу решают лингвистические процессоры, однако при разборе документов в направлении текстового анализа данных (text mining) рассматривается, как правило, отдельный (уже существующий) экземпляр документа, анализ которого позволяет определить ключевые слова для генерации определенных выводов. В технологии лексикологического синтеза одним из этапов предусматривается проведение углубленного анализа комплекса документов определенного вида. На его основе осуществляется классификация информационных потоков, образующих документы анализируемого вида, и синтезируется совокупность опорных слов, объединяемых в лексикологическое дерево. Преимуществом использования лексикологического синтеза является значительное (до 5–7 раз) сокращение трудозатрат пользователей при автоматизированном формировании документов, исключение погрешностей при их создании, а также освобождение пользователей от необходимости изучения правил оформления документов.

Блок-схема последовательности операций, иллюстрирующая сущность способа преобразования слабоформализуемых документов для сокращения объема при хранении, изображена на рис. 1.

Документ создается на базе формы, выбираемой по виду формируемого документа из возможного комплекса (совокупности документов определенного вида) и определяемой соответствующим индексом.

При создании документа его текстовая часть формируется автоматизированным лексикологическим способом путем обхода лексикологического дерева [18–20]. Каждой формулировке документа ставится в соответствие основное слово, выбор которого однозначно определяет наличие конкретной формулировки в документе. Такие слова являются опорными, на их основе составляется лексикологическая схема формируемого документа. Взаимная зависимость опорных слов в совокупности определяет последовательность обхода маршрута формирования документа. На базе предварительного анализа структуры документа выявляются основные разделы, которые должны или могут в нем присутствовать. Условные наименования таких разделов составляют основу синтезируемой совокупности опорных слов.

В рамках каждого зафиксированного раздела документа на этапе предварительного анализа комплекса документов заблаговременно выявляются составные элементы, которые должны или могут входить в состав раздела (слово, фраза, текстовый фрагмент). Для каждого подобного составного элемента определяется опорное слово (или их совокупность), выбор которого в дальнейшем будет однозначно определять внедрение в документ соответствующего компонента.

Если фрагмент текста документа содержит значительное число строк и всегда присутствует в документе в строго определенной последовательности построения предложений, то данный фрагмент текста определяется одним опорным словом. Например, выбор опорного слова «норма» в медицинском протоколе может означать необходимость внедрения в документ целой фразы, характеризующей соответствие описываемого параметра принятым нормам. Однако в большинстве случаев опорные слова соответствуют более коротким формулировкам, присутствие которых в документе необходимо в соответствии с описываемой ситуацией. Поэтому


**Рис. 1** Блок-схема последовательности операций при хранении слабоформализуемых документов

число опорных слов в значительной степени зависит от вариативности документа и требуемой степени гибкости лексикологического дерева, которая определяется детальностью анализа при проведении синтеза совокупности опорных слов. В связи с этим 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.

Поскольку каждой формулировке соответствует опорное слово, выбор такого слова будет предопределять и возможность индексирования выбора. Поэтому в процессе обхода лексикологического дерева при формировании документа осуществляется фиксация документарной индексной последовательности, создаваемой нарастающим итогом путем пошаговой конкатенации индексов опорных слов, в единую кодовую последовательность, соответствующую формируемому документу.

Сформированная документарная индексная последовательность затем подвергается процедуре сжатия информации, для чего может использоваться один из известных методов (например, алгоритм Хаффмана), после чего сохраняется на носителе информации.

При необходимости восстановления документа информационный пакет считывается и подвергается процедуре, обратной сжатию, что позволяет восстановить исходную индексную последовательность. Далее восстанавливается форма документа на основе записанного индекса формы, после чего происходит восстановление содержательной части документа по зафиксированной индексной последовательности и лексикологическому дереву.

## Этап создания документа

На рис. 2 приведена блок-схема последовательности операций, иллюстрирующая собственно процесс автоматизированной фиксации индексной последовательности при формировании документа.

Фиксация индексной последовательности, соответствующей выбираемым опорным словам, осуществляется пошагово в рамках организуемого цикла выбора опорных слов. В случае отсутствия в лексикологическом дереве 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.





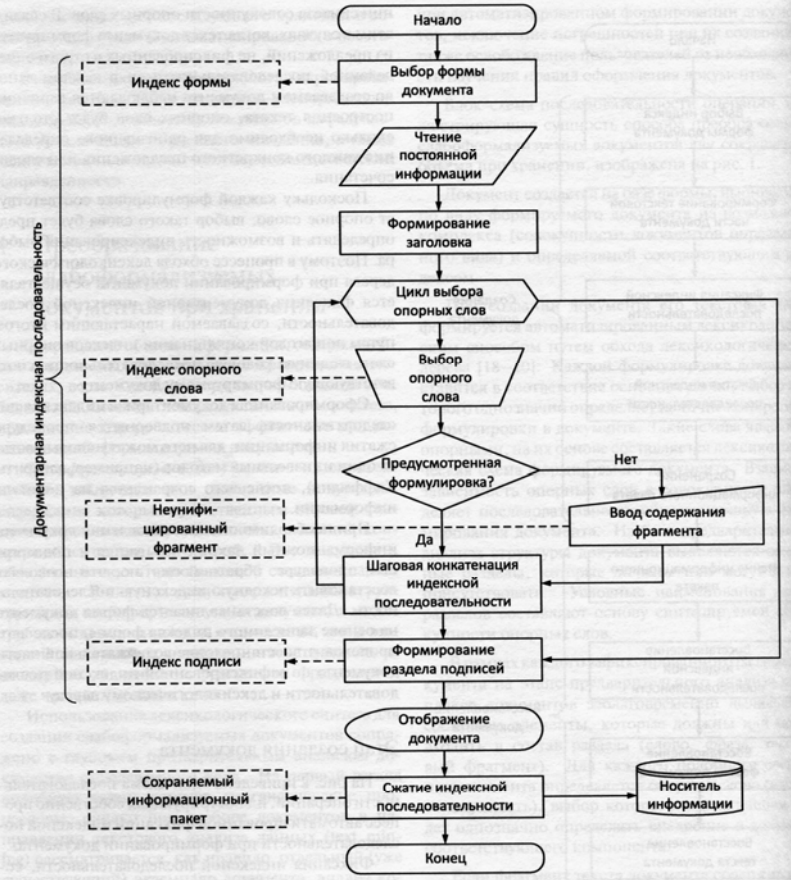


**Рис. 2** Блок-схема операций при фиксации документарной индексной последовательности

Пример фиксации фрагмента индексной последовательности при автоматизированном формировании протокола осмотра пациента в ходе проведения гастроскопии[1], приведен на рис. 3. Первый уровень — тип гастроскопии, второй уточнение типа (причины ургентной гастроскопии, манипуляции при лечебной гастроскопии), третий — конкретизация манипуляций при лечебной гастроскопии, четвертый характеристика состояния пациента, пятый — характеристика оперативных действий.

На лексикологической схеме, например, показано, что при выборе типа гастроскопии можно выбрать лечебную. В этом случае для уровня типа гастроскопии 1 фиксируется индекс 3.

[1]Пример приведен для случая применения рассматриваемого способа в сфере медицины ввиду чрезвычайно высокой вариативности документов, формируемых в этом направлении деятельности, и, следовательно, наибольшей демонстрации эффективности.

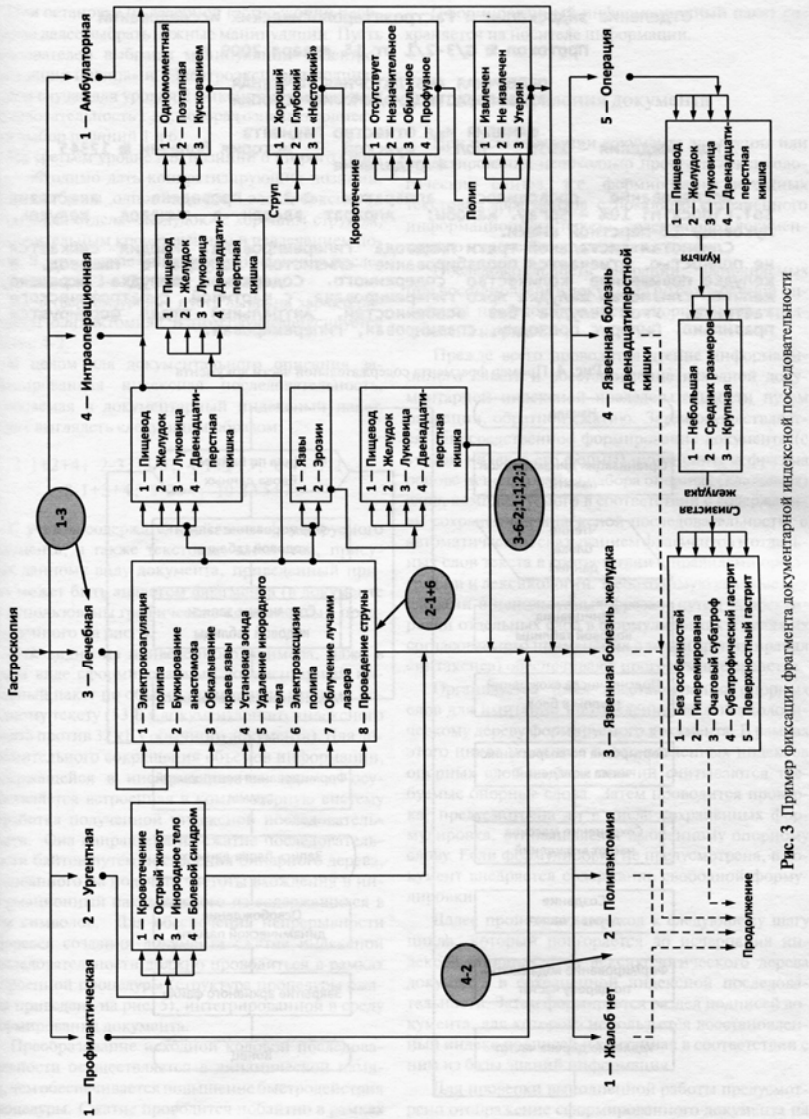


**Рис. 3** Пример фиксации фрагмента документарной индексной последовательности





Отделение эндоскопии и гастроэнтерологических исследований

Протокол № G/3-2/1 от 15 января 2009 г.

ПЕРВИЧНАЯ АМБУЛАТОРНАЯ ЛЕЧЕБНАЯ
ЭЗОФАГОГАСТРОДУОДЕНОФИБРОСКОПИЯ

ФАМИЛИЯ ИМЯ ОТЧЕСТВО Пациента
Год рождения - 1950 Пол - мужской История болезни № 12345
4 отделение

Исследование проводилось аппаратом G-3. Проведена анестезия Sol.Lidocaini 10% - Spray. Жалобы: Аппарат введен в пищевод, желудок, луковицу 12-перстной кишки.

Слизистая дистальной трети пищевода гиперемирована. Кардия смыкается не полностью. Отмечается пролабирование слизистой желудка в пищевод. В желудке повышенное количество содержимого. Содержимое желудка окрашено желчью. Слизистая желудка ярко гиперемирована, с картиной субатрофического гастрита. Угол желудка без особенностей. Антральный отдел формируется правильно. Пилорус проходим, спазмирован, гиперемирован.

**Рис. 4** Пример фрагмента содержательной части документа

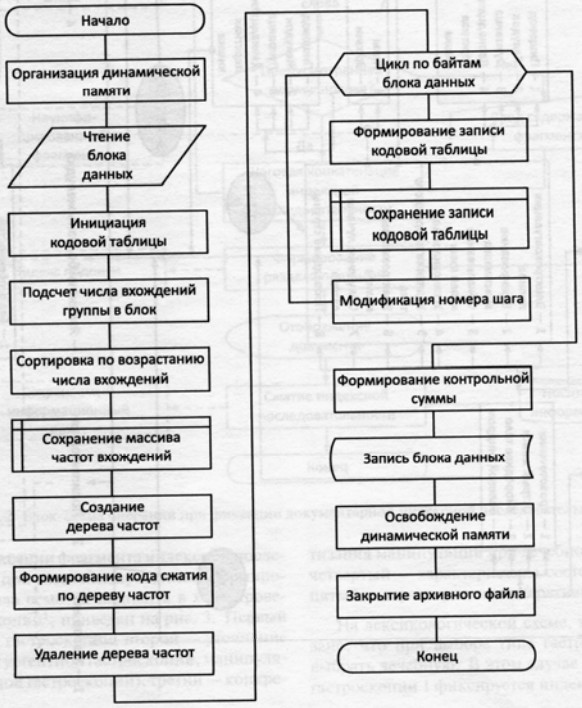


**Рис. 5** Структура процедуры сжатия документарной кодовой последовательности

При остановке на лечебной гастроскопии необходимо далее выбрать нужные манипуляции. Пусть пользователем выбраны манипуляции «Электрокоагуляция полипа» и «Электроэксцизия полипа». В этом случае для уровня 2 фиксируется индексная последовательность 1+6, которая означает совместный выбор позиций 1 и 6.

На третьем уровне для позиции 6 второго уровня необходимо дать конкретизирующие позиции. Пусть выбрана одномоментная электроэксцизия полипа для отдела «Желудок» с хорошим струпом, незначительным кровотечением и извлечением полипа. В этом случае фиксируется индексная последовательность 3-6-2;1,1;2;1.

На четвертом уровне пусть выбрана характеристика «Полипэктомия». В этом случае фиксируется индекс 4-2.

В целом для документального описания зафиксированная индексная последовательность, включаемая в документарный индексный пакет, может выглядеть следующим образом:

1-1+2+4, 2-3, 3-0, 4-1+2+3, 6-1; 7-1;
8-1+3+4, 9-2+4, 10-1+5+7

С учетом содержательной части формируемого документа, а также текстовых элементов, присущих данному виду документа, приведенный пример может быть аналогом фрагмента (в документе не использованы графические компоненты), представленного на рис. 4.

Как видно из приведенного примера, даже в таком виде сформированный документарный индексный пакет по объему значительно уступает исходному тексту (53 Б у документарного индексного пакета против 32 кБ у обычного документа). Для дополнительного сокращения объемов информации, содержащейся в информационном пакете, осуществляется встроенная в компьютерную систему обработка полученной индексной последовательности. Она направлена на сжатие последовательности байтов путем построения бинарного дерева, основанного на подсчете частоты вхождения в информационный пакет каждого из содержащихся в нем символов. Для обеспечения непрерывности процесса создания документа сжатие индексной последовательности должно проводиться в рамках встроенной процедуры (структура процедуры сжатия приведена на рис. 5), интегрированной в среду формирования документа.

Преобразование исходной кодовой последовательности осуществляется в динамической памяти, чем обеспечивается повышение быстродействия процедуры. Сжатие проводится побайтно в рамках считанного блока данных.

Сформированный информационный пакет сохраняется на носителе информации.

## Этап восстановления документа

При восстановлении документа для чтения или редактирования необходимо провести лексикологический синтез, т. е. формирование исходных текстовых фрагментов с помощью сохраненного информационного пакета, содержащего документарную индексную последовательность.

Последовательность операций, выполняемых при восстановлении документа по сохраненной индексной последовательности опорных слов, представлена на рис. 6.

Прежде всего проводится чтение информационного пакета и восстановление исходной документарной индексной последовательности путем операции, обратной сжатию. Затем осуществляется непосредственное формирование документа (с учетом индекса его формы) путем синтеза фраз на основе использования набора опорных (ключевых) слов, комплектуемого в соответствии с содержанием сохраненной индексной последовательности, с автоматическим связыванием фрагментов и отдельных слов текста в соответствии с правилами орфографии и лексикологии. Необходимую связь между словами в используемых фразах (путем корректировки отдельных слов в формулировках в целях их согласованного применения с точки зрения правил синтаксиса) обеспечивают программные средства.

Организуется цикл восстановления опорных слов для имитации прохождения по лексикологическому дереву формируемого документа. В рамках этого цикла на основе восстановленных индексов опорных слов из базы знаний считываются требуемые опорные слова. Затем проводится проверка, предусмотрена ли в числе сохраненных формулировка, относящаяся к выбранному опорному слову. Если формулировка не предусмотрена, в документ внедряется содержание свободной формулировки.

Далее происходит переход к следующему шагу цикла, который повторяется до исчерпания индексов опорных слов лексикологического дерева документа в сохраненной индексной последовательности. Затем формируется раздел подписей документа, для которого используется восстановленный индекс подписей и считанная в соответствии с ним из базы знаний информация.

Для проверки выполненной работы предусмотрено отображение сформированного документа на экране монитора.





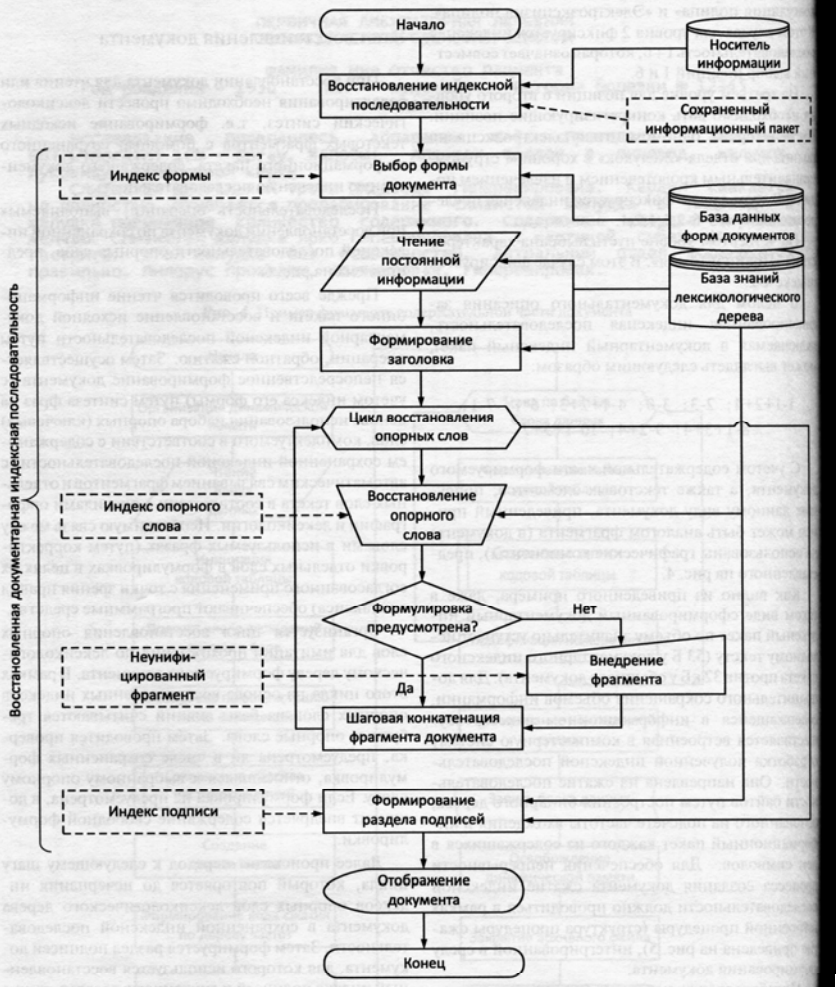


**Рис. 6** Структура процесса восстановления документа

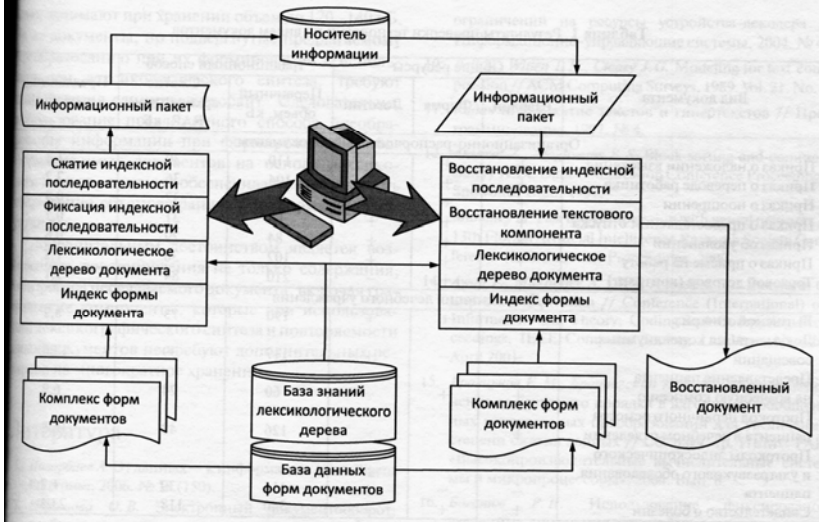


**Рис. 7** Структура системы создания и восстановления слабоформализуемых документов

## 5 Системная реализация комплекса создания и хранения слабоформализуемых документов

Структура системы хранения слабоформализуемых документов приведена на рис. 7

Автоматизированное формирование документа осуществляется с использованием специализированной программы и стандартного компьютера. Формирование документа ведется в диалоговом режиме с автоматическим пошаговым «наращиванием» объема текста за счет внедрения конкретных формулировок, связанных с зафиксированными индексами опорными словами. Унифицированная постоянная информация внедряется в документ автоматически.

При создании документа с помощью его лексикологического дерева, связанного с базой знаний и комплексом форм документов, фиксируется индексная последовательность формируемой информации, которая после дополнительной обработки, направленной на сжатие длины документарной индексной последовательности, сохраняется на носителе информации (например, на жестком диске).

При открытии документа на чтение или редактирование после обработки индексной последовательности, обратной сжатию, осуществляется восстановление индексной последовательности при использовании согласованного лексикологического дерева документа, связанного с комплексом форм документов и базой знаний, содержащей заготовки фрагментов документа.

Восстановление формы документа проводится из базы данных форм документов на основе сохраненного индекса формы, входящего в состав документарной индексной последовательности, после чего постоянная информация считывается из базы знаний и формируется заголовок.

## 6 Экспериментальная проверка предлагаемого метода

Проведена экспериментальная проверка предлагаемого способа организации хранения слабоформализуемых документов, формируемых в интерактивном режиме с использованием технологии лексикологического синтеза. Проверка осуществлялась на примере организационно-распорядительных документов, учетных и отчетных





**Таблица 1** Результаты проверки технологии по видам документов

| Вид документа | Общие ресурсы | | | Традиционный способ | | Предлагаемый способ, кБ |
|---|---|---|---|---|---|---|
| | Шаблон | Форма | Логотип | Первичный объем, кБ | Сжатие архиватором RAR, кБ | |
| **Организационно-распорядительные документы** | | | | | | |
| Приказ о наложении взыскания | + | + | + | 110 | 39 | 1,4 |
| Приказ о переводе работника | + | + | + | 104 | 36 | 2,2 |
| Приказ о поощрении | + | + | + | 95 | 32 | 0,9 |
| Приказ о предоставлении отпуска | + | + | + | 98 | 31 | 0,7 |
| Приказ об увольнении | + | + | + | 84 | 29 | 0,6 |
| Приказ о приеме на работу | + | + | + | 102 | 36 | 1,2 |
| Трудовой договор (контракт) | + | + | – | 210 | 84 | 2,1 |
| **Документация лечебного учреждения** | | | | | | |
| Выписной эпикриз | + | + | + | 190 | 73 | 8,2 |
| Документация консилиумных совещаний | – | – | + | 117 | 39 | 12,6 |
| Представление пациента на врачебную комиссию | + | + | | 60 | 24 | 6,9 |
| Протокол первичного осмотра пациента в лечебном отделении | + | – | – | 126 | 48 | 16,8 |
| Протоколы эндоскопического и ультразвукового обследования пациента | – | – | + | 114 | 39 | 1,7 |
| Свидетельство о болезни | + | + | + | 340 | 118 | 23,4 |
| **Документация высшего учебного заведения** | | | | | | |
| Акт готовности кафедры к новому учебному году | – | + | – | 105 | 49 | 1,3 |
| Отчет о практиках | – | + | – | 159 | 63 | 2,4 |
| Отчет о работе ГЭК | – | + | – | 117 | 41 | 3,6 |
| Отчет о работе кафедры в учебном году | + | + | – | 294 | 92 | 12,7 |
| Протокол заседания ГЭК | + | + | | 92 | 27 | 3,8 |

документов медицинской направленности, а также документации кафедрального уровня высшего учебного заведения.

В соответствии с регламентами, действующими в исследуемых областях деятельности, оригиналы документов исследуемых групп должны храниться в твердых (бумажных) копиях. Их электронные версии, как правило, сохраняются на носителях информации в интересах оперативного обращения к ретроспективной информации.

Традиционным способом создания этих документов на основе установленных форм является использование текстового процессора Microsoft Word. При этом в ряде документов используются заранее подготовленные шаблоны и применяются логотипы организации.

Усредненные результаты экспериментальной проверки предлагаемого способа по видам документов приведены в табл. 1.

## 7 Заключение

Разработана технология хранения слабоформализуемых документов, создаваемых с помощью лексикологического синтеза. Подход апробирован на примерах подготовки к хранению организационно-распорядительных документов, документации лечебного учреждения, а также кафедральных документов высшего учебного заведения. Реализована возможность сокращения объемов сохраняемой информации благодаря формированию сохраняемых индексных последовательностей, содержащих индексы форм документов и их содержательных компонентов. Доказаны преимущества предлагаемого способа преобразования слабоформализуемых документов для минимизации их объема при хранении.

К примеру, деловые документы, содержащие как логотип организации, так и текстовую информа-

цию, занимают при хранении объем до 120–140 кБ. Те же документы, но подвергнутые предлагаемому преобразованию при их формировании с использованием лексикологического синтеза, требуют для хранения единицы килобайт. Следовательно, использование предложенного способа преобразования информации при формировании слабоформализуемых документов на основе лексикологического синтеза обеспечивает возможность сокращения объемов хранимой информации в десятки раз.

Дополнительным достоинством является возможность восстановления не только содержания, но и формы передаваемого документа, включая графические компоненты, которые при использовании лексикографического синтеза и повторяемости видов документов не требуют дополнительных ресурсов на многократное хранение.

## Литература


1. *Назарбаев А.* От данных — к информации // Intelligent Enterprise, 2006. № 18 (150).
2. *Монашова О. В.* Электронный документооборот: проблемы долговременного хранения электронных документов с ЭЦП // VII Международная конференция «Право и Интернет». www.ifap.ru/pi/07
3. *Коупланд Л.* Сжатие данных и изображений // Computerworld, 2000. № 33.
4. *Лобанов А. К.* Методы построения систем хранения данных // Jet Info Online, 2003. № 7
5. *Тойго Д. В.* Сжатие сохраняемой информации набирает обороты // Сети и системы связи, 2006. № 6.
6. *Беляев А. В.* Методы и средства защиты информации. — СПбГТУ, 2000.
7. *Ватолин Д., Ратушняк А., Смирнов М., Юкин В.* Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео. — М.. Диалог-МИФИ, 2003.
8. *Каньковски П.* Делаем из слона муху // Компьютерра, 2003. № 18.
9. *Смирнов М. А.* Использование методов сжатия данных без потерь информации в условиях жестких ограничений на ресурсы устройства-декодера // Информационно-управляющие системы, 2004. № 4.
10. *Bell T., Witten I. H., Cleary J. G.* Modeling for text compression // ACM Computing Surveys, 1989. Vol. 21. No. 4.
11. *Кадач А. В.* Сжатие текстов и гипертекстов // Программирование, 1997. № 4.
12. *Arnavut Z., Magliveras S. S.* Block sorting and compression // IEEE Data Compression Conference Proceedings. Snowbird, Utah, 1997
13. *Mahoney M.* Fast text compression with neural networks // 13th Florida Artificial Intelligence Research Society Conference (International) Proceedings, 2000.
14. *Awan F., Mukherjee A.* LIPT: A lossless text transform to improve compression // Conference (International) on Information and Theory: Coding and Computing Proceedings, IEEE Computer Society. Las Vegas Nevada, April 2001.
15. *Кравцунов Е. М., Брайловский И. В.* Адаптивное изменение алфавитного порядка в алгоритмах обобщенных интервальных преобразований для увеличения степени сжатия данных // Сб. науч. тр. ИВМС РАН «Высокопроизводительные вычислительные системы и микропроцессоры», 2004. Вып. 6.
16. *Богатов Р. Н.* Использование фиксированно-удаленных контекстов для повышения степени сжатия данных // Омский научный вестник, 2006. № 6 (41). — Омск: Изд-во ОмГТУ.
17. *Ахуджа Р.* Сжатие данных в DB2 9. www.ibm.com/developerworks/ru/library/dm-0605ahuja.
18. *Черников Б. В.* Принцип лексикологического синтеза в технологии создания текстовых документов // Секретарское дело, 2000. № 1.
19. *Черников Б. В.* Способ автоматизированного лексикологического синтеза документов. Патент РФ № 2253893, 2005.
20. *Черников Б. В.* Системные аспекты создания слабоформализуемых документов на основе способа лексикологического синтеза // Материалы Всеросс. межвуз. науч.-практ. конф. «Актуальные проблемы информатизации. Развитие информационной инфраструктуры, технологий и систем». М.. МИЭТ. 2007