УДК 658.512.011

ОПТИМИЗАЦИЯ ДОСТОВЕРНОСТИ ИНФОРМАЦИИ НА ОСНОВЕ БАЗЫ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ И ОСОБЕННОСТЕЙ ПРАВИЛ КОНТРОЛЯ БАЗЫ ЗНАНИЙ

Жуманов И. И., Каршиев Х. Б.
h-qarshiyev@samdu.uz;
Самаркандский государственный университет

Сформулирована проблема повышения достоверности информации в системах электронного документооборота на основе методов и алгоритмов, основанных на использовании типичных инструментов поиска, распознавания, классификации документов, а также генерации, трансляции текстов с одного языка в другой, контроля и коррекции орфографических ошибок различной кратности. Предложены концептуальные принципы использования информационной избыточности различной природы – статистической, естественной, структурно-технологической, семантической при повышении достоверности информации. Разработаны алгоритмы, использующие логические, семантические и структурно-технологические связи между элементами документа, получены инструменты использования перекрестных взаимосвязей между отдельными или группами записей в составе кадра информации. Предложены инструментарии контроля элементов электронных документов с опорой на выбранные элементы из базы знаний и на основе использования экспертных систем. Разработаны алгоритмы оптимизации размещения электронных документов со множеством элементов, атрибутов, концептов, фрактальных характеристик в базах данных и базы знаний с механизмом регулирования переменных на основе генетических алгоритмов поиска для выбора объекта с нужными характеристиками. Получены оценки меры близости элементов документа, проведен анализ значений коэффициента выигрыша в достоверности зависимости от объема информации при различных значениях вероятности необнаруженных ошибок по синтезированным алгоритмам. Исследована эффективность реализованной технологической схемы по критерию трудоёмкости обработки информации.

Ключевые слова: электронный документ, достоверность информации, информационная избыточность, метрика близости, поиск, распознавание, классификация, база данных, база знаний, коэффициент расхождения, трудоёмкость обработки информации.

Цитирование: *Жуманов И. И., Каршиев Х. Б.* Оптимизация достоверности информации на основе базы электронных документов и особенностей правил контроля базы знаний // Проблемы вычислительной и прикладной математики. — 2019. — №3(21). — С. 57–74.

1 Актуальность темы

В настоящее время работы в области создания систем электронного документооборота (СЭД) ведутся во многих странах, в том числе в нашей республике достигнуты определенные успехи в этом направлении [1–3].

Технологии СЭД задаются широким спектром приложений, такими как средства чтения электронной почты, редакторы контроля орфографии естественных языков,

инструменты машинного перевода, синтезаторы и анализаторы речи, текстов, обработки служебных документов, программы, способных читать голосом текстовых файлов или текста и др. [4].

В отмеченном направление исследования существуют технологии фирм Microsoft, Lucent, Lernout & Hauspie, Unisys, Elan и др. Применяются пакеты программ корпорации Unisys, которые построены по технологии NLU для распознавания и "понимания" человеческой речи, а также ведения полноценного диалога с компьютерами [2, 3].

Функционируют типовые программные обеспечения СЭД, направленные на решении, в первую очередь, задач учёта, контроля исполнения организационно-распорядительных документов (ОРД), анализа и поиска, также формирования баз данных (БД) и баз знаний (БЗ) [1].

Ключевым вопросом настоящего исследования является повышение достоверности информации в СЭД. Информации искажаются при вводе по вине человека-оператора, из-за погрешности устройства сканирования и распознавания, влияния помех в каналах связи, а также неудовлетворительного обслуживания системы техническими работниками.

Наибольшее количество ошибок, как правило, допускается на этапах сканирования, ввода информации оператором. Использование ошибочной информации, в свою очередь, вызывает значительные искажения работы СЭД и приводит к потере ценности документа, ее информации почти полностью.

Распространенными методами обеспечения требуемой достоверности информации являются корректирующие коды, позволяющие обнаруживать и исправлять ошибки, а также аппаратурные способы помехоустойчивой передачи информации систем обратной и безобратной связи [5].

По мнению многих специалистов из-за дешевизны и простоты применения более эффективны методы компьютерного контроля достоверности информации, к числу которых относятся методы контрольных сумм, контроля по модулям суммирования, подходы, направленные к использованию информационной избыточности различной природы [6–9].

Современные методы, применяемые в СЭД далеки от совершенства, вследствие чего не обеспечивается требуемая точность, полнота, релевантность, достоверность электронных документов (ЭД). Нарушение сохранности, достоверности и целостности документа, содержащейся в нем информации, наиболее вероятно происходит в процессах доставки, хранения, передачи документа, прохождения им этапов согласования и утверждения. Кроме того, традиционные подходы, направленные на обеспечение требуемой достоверности информации ЭД, не оправдывают себя по причинам дороговизны, узкой специализации инструментальных средств [10].

Перспективным и эффективным направлением совершенствования и развития методов обработки информации в СЭД является использование, разработка и реализация алгоритмов, определяющих выполнения типичных функций, в частности поиска, распознавания, классификации, анализа, синтеза, генерации, трансляции текстов с одного языка в другой, а также контроля орфографии, коррекции ошибок различной кратности [11].

Методология повышения достоверности информации, посвященная разработке методов, моделей, алгоритмов и систем недостаточно развита, так что назрела необходимость, требующая более обширного исследования проблемы, а также разработки

программ контроля достоверности информации различной природы, применения для повышения эффективности систем управления [12].

Результаты анализа известных работ, программных технологий и продуктов позволяют спланировать специфические исследования, направленные на решение проблемы повышения достоверности информации в СЭД на принципиально новой научно-методической основе.

2 Базовые подходы и принципы повышения достоверности информации

2.1 Оценка показателя достоверности информации ЭД

Для повышения достоверности информации ЭД предлагается принцип контроля верности элементов, ключевых концептов (слов, фраз, терминов) путем сравнения введенного документа с эталонным – оригиналом, размещенным в БД. Пусть задано множество ЭД $\Omega = \{\omega_1, ..., \omega_r\}$, представляющее последовательность цифровых форматов, либо цифровых изображений образа-документа, которые могут содержать различные искажения информации, например частичная, полная замена слов, строк, абзацев и страниц текста и т.д. Для оценки достоверности информации предлагается коэффициент расхождения введенного и эталонного документов

$$K_{расхожс} = \frac{K_{ввод} - K_{изм}}{K_{ориг}},$$

где $K_{ориг}$ – информационная емкость эталонного оригинала документа; $K_{ввод}$ – информационная емкость введенного документа; $K_{изм}$ – число элементов (концептов) ЭД, отличающихся от оригинала.

Вероятность расхождения оригинала и введенного документа должна приближаться к нулю

$$P_{расхоз} = \lim_{K_{ориг} \to \infty} K_{расхожс} \to 0.$$

Существующая технология повышения достоверности информации рассматривается в следующих вариантах:

– сравнение отсканированных документов с эталонным на ПЭВМ;

– попиксельное сравнение изображения введенного документа с изображением эталонного на основе средств офисной техники.

Однако, эти подходы не обеспечивают достоверности информации до требуемого уровня и отличаются вычислительной сложностью алгоритмов, вследствие чего ограничивается их применение.

Предложен подход, направленный на повышение достоверности информации ЭД в следующих вариантах:

– сопоставление каждого поступившего документа с эталонным, размещенным в БД;

– сопоставление оцифрованного изображения введенного ЭД с изображением оригинала документа, размещенным в БД;

– решение задачи контроля достоверности ЭД, основанного на применении инструментов поиска, распознавания и классификации документа.

2.2 Конструктивные подходы и принципы решения задач

Предметная область характеризуется следующими особенностями:

– наличие большого количества пользователей, решаемых задач, различных требований к достоверности информации;

– передача, хранение и обработка большого объема цифровой, текстовой, алфавитно-цифровой информации разнородной природы;

– необходимость ведения контроля орфографических ошибок, возникающих в работе пользователей и канонической структуры БД;

– возможность использования статистических, логических, семантических, технологических связей между элементами и отношений концептов документов, а также применения физической БД.

Алгоритмы, использующие логические, семантические и структурно-технологические связи между элементами, атрибутами, фреймами ЭД, представляют собой наиболее эффективные, простые инструменты достоверного восстановления информации, главным принципом которых является использование экспертных систем (ЭС), БД, БЗ, а также программных методов контроля достоверности информации [13, 14]. При структуризации, индексации и идентификации входных документов решены следующие задачи:

– анализ особенностей ЭД, целевых функций оптимизации достоверности, разработка упрощенных схем программной реализации, механизмов применения элементов ЭС, БД и БЗ;

– определение состава элементов входных ЭД, событий, порядка и пределов изменения параметров, устойчивых взаимосвязей между ними;

– описание входных потоков с выделением групп информации, описывающих однородные события, каждая из которых представляется обобщенными кадрами информации (формами входных документов). Обобщенный кадр содержит совокупность записей (параметров, полей, реквизитов);

– анализ структуры записей, входящих в состав каждой группы входной информации (обобщенных кадров входной информации), в частности уточняется смысловое значение каждой записи, устанавливается отношение каждой записи или группы связанных между собой записей;

– анализ каждого типа записей входной информации (показателя, реквизита и т.п.), установление существующих закономерностей, описание данного показателя, определение пределов изменения, запрещенные значения (например, граничные значения для конкретной предметной области применения создаваемой системы);

– установление наличия формализованных БЗ по данному типу записей (таблицы численных значений, словари символьных записей, классификаторы, справочники и др.);

– анализ перекрестных взаимосвязей между отдельными записями или группами записей в составе кадра информации;

– выявление устойчивых смысловых и логических связей между записями;

– формализация семантических и логических условий (критериев) для контроля достоверности формируемых входных кадров;

– построение алгоритмов первичной обработки входной информации;

– определение шагов алгоритма, в которых целесообразно включать элементы БЗ и ЭС;

– выработка решающих правил для каждого шага итерации алгоритмов обработки информации;

– синтез функциональной (логической) структуры СЭД с включением ЭС, БД и БЗ.

2.3 Принципы повышения достоверности информации на основе применения БЗ

На входе формируется несколько десятков типов и разновидностей ОРД в зависимости от функциональных возможностей СЭД [15, 16].

Входная информация структурирована в виде потоков входных документов. Каждый ОРД, включает основные характеристики концептов: реквизиты, первичные идентификационные данные, номер и дата документа, вид документа, адресаты создателя документа, содержание документа.

Предварительным анализом каждого документа, групп реквизитов устанавливаются характерные их особенности, например, термины делопроизводства, указатели на справочные данные, которые позволяют их однозначно идентифицировать. Дата документа представляет собой цифровую, строго структурированную запись и др.

Установлено, что использование методов повышения достоверности информации могут выполняться путем сопоставления со справочниками, представляющими собой элементы БЗ соответствующего направления. В качестве элементов БЗ могут использоваться также справочники обозначений к функциональным возможностям системы.

На следующем этапе совершенствования методов требуется более глубокий анализ и использование специальных знаний и экспертных оценок, выявление устойчивых перекрестных связей, формирование на их основе логические и семантические критерии контроля.

Вышесказанное иллюстрируются использованием результатов следующих примеров:

– дата документа не может быть меньше по значению даты основания для документа;

– если таковой предусмотрен, то дата оформления и его регистрации не может быть меньше даты документа;

– если использовать цифру, обозначающую месяц, то ее значение не может быть больше 12;

– максимальная дата должна удовлетворять ограничению исходя из номера месяца (с учетом максимальной даты февраля по високосным и невисокосным годам).

Программы правил контроля с опорой на выбранные элементы БЗ являются реализациями механизмов использования ЭС. Причем, принцип действия элементов ЭС состоит в сопоставлении подготовленных данных с информацией из встроенной БЗ с критериями семантического и логического контроля и выработки реакции соответствующего звена.

2.4 Принципы повышения достоверности информации на основе применения ЭС

Можно выделить два вида реакции ЭС на обнаружение несоответствия входной информации и критериев, правил контроля в БЗ:

– сигнализация об обнаруженных несоответствиях (подсказка на необходимость дополнительного анализа отмеченной входной информации) с возможностью прохождения «дефектной» входной информации на верхние уровни системы;

– блокировка дальнейшего прохождения входной информации на следующие уровни до устранения противоречий между входной информацией и критериями из состава встроенной БЗ.

В первом случае, возрастает поток недостоверной информации, проникающей на следующие уровни системы. Во втором случае, снижается оперативность работы си-

стемы из-за инерции, связанной с выяснением и устранением возникающих ситуаций. В СЭД целесообразно использование сочетания обоих вариантов.

В частности, элементы ЭС распределены по всем иерархическим уровням системы, включая верхний уровень. При этом ЭС верхнего уровня предназначена для функционирования по второму варианту, отсекая попадание недостоверной информации на основе использования БД.

ЭС в нижнем уровне иерархии системы функционирует по первому варианту. При этом минимизируется задержка входной информации на нижних звеньях и обеспечивается поэтапное повышение достоверности при прохождении входной информации через контроль.

3 Алгоритмы повышения достоверности информации на основе использования перекрестных, логических и семантических связей элементов документа

3.1 Принципы повышения достоверности информации на основе логических и семантических связей элементов документа

Существующие логические и семантические связи между элементами документа, например «Контрольная карточка» позволили разработать более простые, но эффективные алгоритмы контроля достоверности информации. Опишем на примерах и проанализируем их эффективность [17, 18].

Входные номера контрольных карточек нумеруются последовательно, что дает возможность при записи этих карточек в БД сравнить их входные номера с номерами предыдущих карточек.

В этих условиях алгоритм контроля запишется так, что входной номер документа будет считаться принятым:

– правильно, если $N_l - N_{l-1} = 1$;

– ошибочно, если $N_l - N_{l-1} \neq 1$, где N_l – номер l-й карточки.

Методом не обнаруживаются ошибки такого типа, когда искажение происходит в обоих номерах, но их разница равняется единице. Тогда вероятность необнаруженных ошибок будет равна [19]

$$P_{HO}^{(i)} = \sum_{i=1}^{m} \left[\frac{C_m^i P^i (1-P)^{m-i}}{n^i \left(A_n^i + n\right)} \right]^2,$$

где $C_m^i = \frac{m!}{i!(m-i)!}$, i – число искаженных цифр в одном номере; m – количество цифр в номере; P – вероятность искажения одной цифры; n – порядок системы счисления (в данном случае $n = 10$); $A_n^i + n$ – всевозможные комбинации i-х цифр.

По аналогичному принципу могут быть проконтролированы даты поступления документов на основе даты отправления. Разница между датами почти достигает величины k. Данная статистика учитывается в алгоритме.

Сообщения по алгоритму считаются принятыми: правильно, если $D_{пос} - D_{отп} \leqslant k$; ошибочно, если $D_{\text{пос}} - D_{\text{отп}} > k$, где $D_{\text{пос}}$, $D_{\text{отп}}$ – соответственно даты поступления и отправления документов.

Алгоритмом не обнаруживаются ошибки, когда искажаются обе даты, но их разность не превышает k. Вероятность необнаруженных ошибок при этом составляет [20]

$$P_{HO}^{(j)} = \sum_{j=1}^{r} \left[\frac{C_m^j P^j (1-P)^{r-j}}{n^j \left(A_n^j + n\right)} \right]^2,$$

где r – количество цифр в датах.

Логическая связь между шифрами руководителя и содержанием документа также позволяет вести контроль.

Рассмотрим пример использования семантических связей между элементами документа. Пусть задано множество шифров руководителя:

$$\Phi = \{\alpha_1, \alpha_2, ..., \alpha_i, ..., \alpha_n\},$$

где α_i – шифр i-го руководителя.

Каждому шифру соответствует определенное подмножество слов и фраз, характеризующее деятельность руководителя (например, для ответственного работника организации по строительству характерны слова: “строительство”, “стройка”, “строить” и т. д.). Такое множество обозначим через φ:

$$\varphi = \varphi\{B_1, B_2, ..., B_i, ...B_n\},$$

где B_i – подмножество слов и фраз, соответствующих шифру α_i.

В качестве контролирующего множества выбираем множество M, определяемое содержанием документа, что позволяет вводить алгоритм контроля достоверности информации по смыслу концептов документа.

Информация x_i считается правильной, если $\beta_i \in M$, и ошибочной, если $\beta_i \notin M$, где β_i – произвольное слово из подмножества B_i.

Алгоритмом не обнаруживаются ошибки в следующих ситуациях: когда шифр α_i искажается на шифр α_j, где $i = 1 \div N$, $j = 1 \div N$, $i \neq j$, а также когда слово β_i искажается на слово β_j, соответствующее шифру α_j.

Вероятность первого события (если шифр состоит из S цифр) принимает вид:

$$P_1 = \sum_{k=1}^{S} \frac{C_S^k P^k (1-P)^{S-k}}{n^k \left(A_n^k + n\right)},$$

где – основание системы, а вероятность второго события (если β_iи β_j состоят из букв) задаётся, как:

$$P_2 = \sum_{l=1}^{R} \frac{C_R^l P^l (1-P)^{R-l}}{n^l \left(A_n^l + n\right)},$$

где R – количество букв в алфавите.

Общая вероятность необнаруженных ошибок алгоритма равна

$$P_{HO} = \sum_{k=1}^{S} \frac{C_S^k P^k (1-P)^{S-k}}{n^k \left(A_n^k + n\right)} \sum_{l=1}^{R} \frac{C_R^l P^l (1-P)^{R-l}}{n^l \left(A_n^l + n\right)},$$

Таким образом, на трех конкретных примерах доказана возможность использования статистических, логических и семантических свойств элементов документа, применяемых в СЭД. Эффективность изложенных алгоритмов повышения достоверности информации достаточно высока.

3.2 Принцип повышения достоверности информации на основе перекрестных связей элементов документа

Пусть оператор создает документ «Приказ» о приеме на работу. Основанием для данного приказа является документ «Заявление». В данном случае проверяются несколько параметров на основании логических критериев и встроенных БЗ.

Контроль достоверности информации документа осуществляется на основании встроенного критерия в БЗ в виде перекрестной ссылки. Алгоритм представляется в следующих шагах.

Шаг 1. Проверяется вид документа «Приказ».

Шаг 2. Находится соответствующий вид документа из справочника перекрестных связей между документами.

Шаг 3. Определяется документ, который является основанием.

Шаг 4. Сравнивается выбранный вид документа с видом документа из БД.

Шаг 5. Если типы совпадают, тогда регистрируется документ.

Шаг 6. Иначе выдается соответствующее сообщение.

Принцип контроля достоверности информации заключается в следующем. При сравнении основания документа «Приказ» необходимо учесть, что адресат в документе «Заявление» и создатель в документе «Приказ» должны совпадать. Это достигается за счет перекрестных связей между документами и связей этих документов со справочником ответственных лиц данного предприятия. В данном случае также проверяется соответствие прав адресата «Заявления» и создателя «Приказа» на принятие и формирование данных видов документов. Для каждого уровня определен уровень доступа к чтению и редактированию определенных документов.

Для решения этой задачи все документы помечаются на редактирование и чтение уровнями доступа. Например, если документ «Приказ» на чтение помечен уровнем 9, а на редактирование уровнем 7, то это означает, что должности до 7 уровня включительно могут читать и редактировать документ, а должности с 8 по 9 уровень могут только читать документ. Должности уровня выше 9 не будут иметь доступ к документу. Функция ЭС направлена на учет взаимосвязи элементов документа, как контролирующая достоверности и правомочности регистрации документов.

Данная функция иллюстрирует последовательное введение и регистрацию документов. Если не зарегистрирован документ «Заявление» на работу, то не может быть зарегистрирован документ «Приказ» на заявление. Еще одним подходом, направленным на повышение качества данной функции является контроль элементов ЭД, таких, как ФИО и должность ответственного лица. Реализованный алгоритм выполняет две функции.

На вход подается дата. Проверяется структура записи даты. Если структура записи совпадает с разрешенными структурами, то выдается 1, иначе выдается 0;

– сравнение дат на меньшее. На вход подаются две даты - дата1 и дата2. Если дата1<дата2, тогда возвращается 1, иначе возвращается 0;

– сравнение дат на большее. На вход подаются две даты - дата1 и дата2. Если дата1>дата2, тогда возвращается 1, иначе возвращается 0;

– сравнение дат на равенство. На вход подаются две даты - дата1 и дата2. Если дата1=дата2 , тогда возвращается 1, иначе возвращается 0;

– сравнение периода даты. В системе определено число в месяцах, показывающее охват периода использования дат от текущей даты назад и вперед. На вход подается дата. Если дата отличается от текущей на большее количество месяцев, чем заданное число в системе, тогда выдается 0, иначе 1;

– день недели. На вход подается дата. По дате определяется день недели. Если день недели входит в нерабочий день или праздник, то выдается 0, иначе 1;

– календарь, которая реализована в виде отдельного модуля и используется для правильного ввода структуры записи даты.

Модифицированный алгоритм представляется в следующих шагах.

Шаг 1. Проверяется вид документа «Приказ».

Шаг 2 . Находится соответствующий вид документа из справочника перекрестных связей между документами.

Шаг 3. Определяется документ, который является основанием.

Шаг 4. Сравнивается вид документа и вид из справочника базы знаний.

Шаг 5. Если типы не совпадают, тогда выдается сообщение.

Шаг 6. Иначе проверяется адресат «Заявление» и создатель «Приказа».

Шаг 7. Если не выполняется условие в шаге 6, тогда выдается соответствующее сообщение.

Шаг 8. Иначе проверяются даты документов «Заявление» и «Приказ».

Шаг 9. Если дата «Заявление» меньше, чем дата «Приказ» регистрируем.

Шаг 10. Иначе выдается соответствующее сообщение.

4 Оптимизация размещения документов в базах данных и базах знаний

Постановка задачи. Опыт зарубежных и отечественных исследований показывает, что для оптимизации размещения данных, большого количества ЭД разнообразных форматов, кадров изображений документов при решении задач поиска, распознавания, классификации, повышения достоверности информации можно успешно применять известные модели и алгоритмы задач оптимального «раскроя», в которых производится поиск локального и глобального экстремума в большом пространстве. При этом эффективные решения таких задач невозможно получить другими методами [21]. Следовательно, проектирование алгоритмов оптимизации размещения объектов – ЭД со множеством элементов, атрибутов, концептов, фрактальных характеристик и др в БД и БЗ, которые эффективно используются при повышение достоверности информации, является остро востребованным.

4.1 Принцип размещения документов на основе прямоугольного раскроя

Задача раскроя любого вида является NP-полной задачей, время решения которой предполагается провести в порядке десятка секунд, что обуславливает модификацию технологии проектирования на основе применения эволюционного моделирования, в частности генетических алгоритмов оптимизации поиска для выбора документа с нужными характеристиками [22]. Подчеркнем особенности предложенного алгоритма решения задачи.

Пусть заданное поле – бесконечная полоса с шириной W , на котором размещаются m прямоугольных предметов $(l_i, w_i), i = 1, 2, ..., m$ – длина и ширина формата документа, который задаётся первоначально.

Ставятся следующие условия:

– (x_i, y_i) – координаты левого нижнего угла прямоугольника на i полосе;

– при размещении на полосе никакие два предмета не пересекаются друг с другом, т.е $((x_i \geqslant x_j + l_j) \vee (x_j \geqslant x_i + l_i) \vee (y_i \geqslant y_j + w_j) \vee (y_j \geqslant y_i + w_i))$ для $i, j = 1, 2, ..., m, i \neq j$;

– никакой предмет не пересекает границы полосы: $(x_i \geqslant 0) \wedge (y_i \geqslant 0) \wedge (y_i + w_i \leqslant W)$ для $i = 1, 2, ..., m$.

Требуется разместить на бесконечной полосе набор прямоугольных предметов без перекрытий так, чтобы занятая ими часть полосы была минимальной по длине.

Требуется найти такой набор (x_i, y_i), чтобы $L = \max(x_i + l_i) \to \min$.

Геометрический смысл переменных $W, L(l_i, w_i), (x_i, y_i), i = 1, 2, ..., m$ изложен в [22].

4.2 Принцип размещения документов на основе раскроя круглых предметов

Теперь рассмотрим задачу размещения на бесконечной полосе с шириной в W, m круглых предметов, радиусы которых известны как r_i.

Ставятся следующие условия:

– (x_i, y_i) – координаты центра окружности i на полосе;

– при размещении на полосе никакие два предмета не пересекаются друг с другом $(x_i - x_j)^2 + (y_i - y_j)^2 \geqslant (r_i + r_j)^2$, для : $i, j = 1, 2, ..., m, i \neq j$;

– никакой предмет не пересекает границ полосы $(x_i - r_j \geqslant 0) \wedge (y_i - r_i \geqslant 0) \wedge (y_i + + r_i \leqslant W)$.

Требуется разместить на бесконечной полосе набор круглых предметов без перекрытий так, чтобы занятая ими часть полосы была минимальна по длине. Правило поиска алгоритма являются эвристическим и содержательно строится так, чтобы размещение было по возможности плотным. В качестве критерия оптимизации рассматривается длина полосы, занятая размещением предметов.

4.3 Принцип оптимального раскроя на основе генетических алгоритмов

Решение задачи раскроя для оптимизации размещения документов в БЗ проводится на основе генетических алгоритмов (ГА), с помощью которых удобно хранить информацию о размещении объектов и располагать рационально каждый следующий объект, учитывая ранее образовавшиеся пустоты [23].

Пустоты учитываются и нумеруются, начиная с числа на единицу большего количества предметов, но со знаком минус.

Одним из особенностей применяемого ГА является блочное декодирование, которое заключается в том, что размещая очередной предмет, он сначала проверяет все образовавшиеся ранее пустоты для достижения большей плотности размещения. ГА использует процедуру декодирования как параметр. Край размещения инициализируется последовательностью вида: $\{Lt, W, Lb\}$ – верх, торец, низ полосы соответственно.

Эволюционная схема ГА, смены поколений, структуры хромосом, адаптация операторов мутации, скрещивания, селекции не чувствительны к геометрии размещаемых предметов. Например, оператор мутации случайным образом переставляет два номера предметов в хромосоме. Оператор скрещивания двух хромосом порождает двух потомков.

Начальный участок дочерней хромосомы совпадает со случайным участком хромосомы одного родителя, а конец состоит из оставшихся предметов, перечисленных в том порядке, в котором они следуют в хромосоме другого из родителей.

ГА, допускающий смену декодеров, и оба декодера были запрограммированы в среде приложений Delphi.

5 Алгоритм повышения достоверности информации на основе априорного словаря признаков

5.1 Методика оптимизации параметров алгоритма

Пусть принимается множество возможных решений $L = \{l_1, ..., l_k\}$ о достоверности информации, которое характеризуется степенью соответствия введенного документа к его эталонному - оригиналу.

Контроль достоверности информации задаётся следующими описаниями:

– исходным множеством последовательности цифровых форматов документов $\Omega = \{\omega_1, ..., \omega_r\}$;

– множеством возможных решений о достоверности информации $L = \{l_1, ..., l_k\}$;

– априорным словарем атрибут-признаков $\mathrm{x_a} = \{x_1, ..., x_N\}$;

– мерой близости элементов документа;

– значением выигрыша за счет контроля достоверности информации;

– расходом времени, затрачиваемого на обработку информации;

– формализованной структурной компонентой ЭД в виде: количества строк N; номеров неполных строк $N_{неполстрок} = |n_{неполстрок1}, ..., n_{неполстрокi}|$; элементов вектора $1 \leqslant n_{неполстрок} \leqslant N_{строк}$, соответствующих порядковому номеру неполной строки (первая и последняя строка абзаца);

– количеством слов в каждой строке $N_{слов} = |n_{слов1}, ..., n_{словi}|$, где $n_{словi}$ – элемент вектора, равный числу в i-й строке документа;

– расположением коротких слов, отражаемых в виде матрицы

$$\|P_{корот.слов}\| = \begin{pmatrix} p_{11} & p_{12} & ..... & p_{1N_{строк}} \\ p_{21} & p_{22} & ..... & p_{2N_{строк}} \\ ..... & ..... & ..... & ..... \\ p_{m1} & p_{m2} & ..... & p_{mN_{строк}} \end{pmatrix},$$

– вектором элементов документа, который задаётся как:

$$P_{корот.слов} = \begin{cases} 1, & j - слов \quad i - короткая\ строка; \\ 0, & в\ противном\ случае, \end{cases}$$

где $i = 1, \ldots, N$; $j = 1, \ldots, m$; m – число слов в строке;

– матрицей площади слов в ЭД, которая отражается в виде

$$\|S\| = \begin{pmatrix} s_{11} & s_{12} & ..... & s_{1N_{строк}} \\ s_{21} & s_{22} & ..... & s_{2N_{строк}} \\ ..... & ..... & ..... & ..... \\ s_{m1} & s_{m2} & ..... & s_{mN_{строк}} \end{pmatrix},$$

где s_{ij} – площадь j-го пикселя, в слове $j = 1, \ldots, m$, i-ой строки $i = 1, \ldots, N_{строк}$;

– матрицей расстояния между словами, задаваемой в виде

$$\|L\| = \begin{pmatrix} l_{11} & l_{12} & ..... & l_{1N_{строк}} \\ l_{21} & l_{22} & ..... & l_{2N_{строк}} \\ ..... & ..... & ..... & ..... \\ l_{m1} & l_{m2} & ..... & l_{mN_{строк}} \end{pmatrix},$$

где $l_{ij} = \sqrt{(x_{ij} - x_{11})^2 + (y_{ij} - y_{11})^2}$ – элемент матрицы;

– x_{ij} и y_{ij} – соответственно, горизонтальным и вертикальным координатами центра масс j-го слова в i-ой строке;

– матрицей, в которой количество букв задаётся по строкам в виде

$$N_{отв} = \left|n_{отв1}, ..., n_{отвN_{строк}}\right|,$$

где $n_{отвi}$ – элемент вектора, равный числу букв i-й строки;

– матрицей расстояния между буквами, задаваемой в виде

$$\|L\| = \begin{pmatrix} l_{отв11} & l_{отв12} & ..... & l_{отв1N_{строк}} \\ l_{отв21} & l_{отв22} & ..... & l_{отв2N_{строк}} \\ ..... & ..... & ..... & ..... \\ l_{отв1} & l_{отв2} & ..... & lN_{строк} \end{pmatrix},$$

где $l_{отвij} = \sqrt{(x_{отвij} - x_{отв11})^2 + (y_{отвij} - y_{отв11})^2}$ – элемент матрицы; $x_{отвjj}$ и $y_{отвjj}$ – соответственно, горизонтальный и вертикальный координаты центра масс j-го отверстия в i-ой строке; о – максимальное число отверстий в строке ЭД;

– вектором, который задаётся числом букв в виде вертикальных линий

$$N_{верт} = \left| n_{верт1}, ..., n_{вертN_{строка}} \right|,$$

где $n_{верт}$ – элемент вектора, равный числу вертикальных линий в словах i-ой строки;

– матрицей расстояния между вертикальными линиями, которая задаётся буквами

$$\|L_{верт}\| = \begin{pmatrix} l_{верт11} & l_{верт12} & ..... & l_{верт1N_{строк}} \\ l_{верт21} & l_{верт22} & ..... & l_{vert2N_{strok}} \\ ..... & ..... & ..... & ..... \\ l_{верт1} & l_{верт2} & ..... & l_{вертN_{строк}} \end{pmatrix},$$

где $l_{вертij} = \sqrt{(x_{вертij} - x_{верт11})^2 + (y_{вертij} - y_{вертt11})^2}$ – элемент матрицы; $x_{вертij}$ и $y_{вертij}$ – соответственно, горизонтальная линия координаты центра масс j-го $j = 1, \ldots, v$, а также вертикальная линия в i-ой строке; v – максимальное число вертикальных линий в строке.

Близость введенного и эталонного документа задаётся в виде:

$$d^2(\omega, \omega_1) = (N^{(p,k)}_{строк} - N^{(q,l)}_{строк})^2 + \sum_{j=1}^{t} (n^{(p,k)}_{неполстрокj} - n^{(q,l)}_{неполстрокj})^2 +$$

$$+ \sum_{j=1}^{N_{строк}} (n^{(p,k)}_{словj} - n^{(q,l)}_{словj})^2 + \sum_{j=1}^{N_{строк}} \sum_{j=1}^{m} (p^{(p,k)}_{коротсловj,i} - p^{(q,l)}_{коротсловj,i})^2 +$$

$$+ \sum_{j=1}^{N_{строк}} \sum_{i=1}^{m} (s^{(p,k)}_{словj,i} - s^{(q,l)}_{j,i})^2 + \sum_{j=1}^{N_{строк}} \sum_{i=1}^{m} (l^{(p,k)}_{словj,i} - l^{(q,l)}_{словj,i})^2 + \sum_{j=1}^{N} (n^{(p,k)}_{отвj} - n^{(q,l)}_{отвj})^2 +$$

$$\sum_{j=1}^{N_{строк}} \sum_{i=1}^{m} (n^{(p,k)}_{отвj,i} - n^{(q,l)}_{отвj,i})^2 + \sum_{j=1}^{N_{строк}} (n^{(p,k)}_{вертj} - n^{(q,l)}_{вертj})^2$$

$$+ \sum_{j=1}^{N_{строк}} \sum_{j=1}^{m} (l^{(p,k)}_{вертиj,i} - l^{(q,l)}_{вертj,i})^2 = d_1^2(\omega_{pk}, \omega_{ql}) + d_2^2(\omega_{pk}, \omega_{ql}) +$$

$$+ d_3^2(\omega_{pk}, \omega_{ql}) + d_4^2(\omega_{pk}, \omega_{ql}) + d_5^2(\omega_{pk}, \omega_{ql}) + d_6^2(\omega_{pk}, \omega_{ql}) + d_7^2(\omega_{pk}, \omega_{ql}) +$$

$$+ d_8^2(\omega_{pk}, \omega_{ql}) + d_9^2(\omega_{pk}, \omega_{ql}) + d_{10}^2(\omega_{pk}, \omega_{ql})$$

где $d_j^2(\omega, \omega_1)$, $j = 1, \ldots, 10$ – эвклидова мера близости параметров:

$$N_{строк}, N_{неполнстрок}, N_{слов}, коротслов, S_{slov}, L_{слов}, N_{otv}, L_{otv}, N_{верт}], L_{верт}.$$

5.2 Анализ достоверности информации на основе меры близости элементов документа

Для упрощенного анализа меры близости элементов с учетом объема обрабатываемой информации представляется следующее соотношение

$$d^2(\omega_{pk}, \omega_{ql}) = \lambda_1 d_1^2(\omega_{pk}, \omega_{ql}) + \lambda_2 d_2^2(\omega_{pk}, \omega_{ql}) + \lambda_3 d_3^2(\omega_{pk}, \omega_{ql}) + \lambda_4 d_4^2(\omega_{pk}, \omega_{ql}) +$$

$$+\lambda_5 d_5^2(\omega_{pk}, \omega_{ql} + \lambda_6 d_6^2(\omega_{pk}, \omega_{ql}) + \lambda_7 d_7^2(\omega_{pk}, \omega_{ql}) + \lambda_8 d_8^2(\omega_{pk}, \omega_{ql}) + \lambda_9 d_9^2(\omega_{pk}, \omega_{ql}) +$$

$$+\lambda_{10} d_{10}^2(\omega_{pk}, \omega_{ql}),$$

где λ_i – параметр, который задаётся N-мерным вектором $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ и характеризует объем информации.

При большом числе измерений меры близости между элементами документа и когда требуется моделирование правила контроля достоверности информации приходится определять пределы допустимых границ, либо применять критерий согласия Колмогорова-Смирнова при функции нормального закона распределения.

Коэффициент расхождения результатов анализа задаётся следующей регрессионной зависимостью

$$K = \lambda_1 d_1^2(\omega_1, \omega_2) + \lambda_2 d_2^2(\omega_1, \omega_3) + \lambda_3 d_3^2(\omega_1, \omega_4) + \lambda_4 d_4^2(\omega_1, \omega_5) + +\lambda_5 d_5^2(\omega_1, \omega_6 +$$

$$+\lambda_6 d_6^2(\omega_1, \omega_7) + \lambda_7 d_7^2(\omega_1, \omega_8) + \lambda_8 d_8^2(\omega_1, \omega_9) + \lambda_9 d_9^2(\omega_1, \omega_{10}) + +\lambda_{10} d_{10}^2(\omega_1, \omega_{11}).$$

Упрощённая модель анализа задаётся в виде

$$K_{схож} = \lambda_0 + \sum_{j=1}^{9} \lambda_j d_j^2(\omega_1, \omega_j).$$

5.3 Реализации технологической схемы повышения достоверности информации

Аппаратная часть технологии повышения достоверности информации основывается на применении традиционных средств офисной техники, локальной вычислительной сети и БД. Программная часть реализована в виде двухуровневой модели повышения достоверности информации.

На первом уровне модели реализованы механизмы занесения оцифрованного изображения документа, его признаков в БД, а на втором уровне реализованы механизмы контроля достоверности информации ЭД.

Точность обработки информации на выходе технологии оценивается по выражению средней погрешности контроля

$$\delta = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{|\Delta_i|}{y_i} \right) \cdot 100\%,$$

где $|\Delta_i|$ – разность элементов, соответственно, введенного и эталонного документа; y_i – задаваемое ограничение на значение разности $|\Delta_i|$; n – число элементов либо концептов документа.

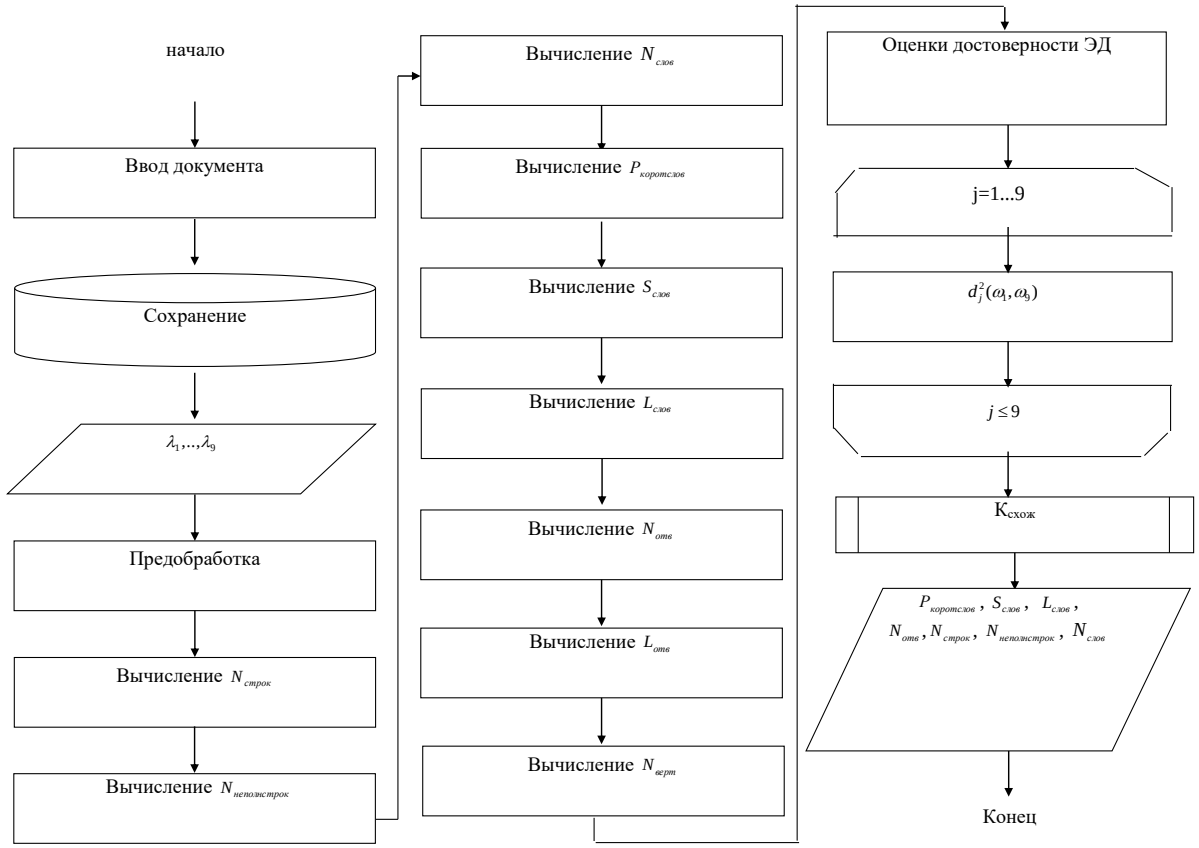

Рис. 1 Схема контроля достоверности информации ЭД

На рис. 1 приведена предложенная технологическая схема для реализации алгоритмов повышения достоверности информации ЭД.

В табл.1 приведены результаты оценки эффективности технологии по критерию трудоёмкости обработки информации в (оп/сек).

5.4 Тестирование технологии повышения достоверности информации

Тестирование проведено на примере 12 документов, элементы которых содержат условные информационные искажения (ошибки).

Для оценки точности обработки информации заданы следующие условия: $\sigma \leqslant 5\%$; время ввода всех документов – 25 с; время сканирования, распознавания и поиска документа на Core 2 Duo 3,3 ГГц/4Гб – 2 с; время сканирования, поиска и распознавания одного документа на Pentium 4/1.8 ГГц/512МБ – 2,5 с; допустимое время T_0 = 1 мин.

Установлено, что благодаря применению предложенной технологии повышения достоверности информации общее время обработки документа уменьшается на 9-10%. Применяющиеся в СЭД методы не обеспечивают достоверности информации до требуемого уровня. Достоверность информации ЭД за счет реализованной технологии повышается выше, чем на 8%. Трудоёмкость и стоимость обработки информации по сравнению с существующей технологией уменьшается в 1,5-2 раза.

Рекомендовано применение алгоритмов повышения достоверности информации, в которых используются:

– правила контроля элемента цифрового изображения введенного и эталонного документа путем сравнения модальными характеристиками;

– словари структурных признаков, рабочий словарь признаков;

– учитываются статистические, логические и семантические связи между элементами и отношения концептов документа;

– эффективные инструменты технологии обработки информации ЭД.

Таблица 1 Эффективность технологической схемы обработки информации ЭД

Документы	Microsoft WORD			Реализованная схема	
	Время поиска, распознавания, классификации, (в сек)	Время контроля достоверности информации (в сек)	Общее время обработки информации (в сек)	Время контроля достоверности информации (в сек)	Общее время обработки информации (в сек)
1	2,5	20,5	48	19, 0	44,12
2	2,5	22,5	50	23,7	45,37
3	2,5	18,5	46	21,4	45,14
4	2,5	20,5	48	18,3	44,83
5	2,5	14,5	42	13,0	37,94
6	2,5	15,5	43	12,5	37,53
7	2,5	16	43,5	12,8	37,78
8	2,5	16,5	44	12,4	37,48
9	2,5	16	43,5	14,0	38,78
10	2,5	16	43,5	15,0	38,91
11	2,5	15,5	43	17,0	38,78
12	2,5	16	43,5	13,8	38,85
Среднее время:		17,3	44,83	13,23	40,46

6 Заключение

Результаты исследований позволяют заключить следующие:

– разработаны логический и физический архитектурные уровни СЭД, определены сущности ЭД, связи между элементами и отношения концептов на основе методик «один-к-одному», «один-ко-многим», «многие-ко-одному» и «многие-ко-многим», а также определены механизмы извлечения специфических характеристик, текстурных свойств и особенностей документов;

– обоснована эффективность использования информационной избыточности различной природы, в частности статистической, естественной, структурно-технологической, семантической, а также концептуальных принципов, методов, алгоритмов и программных комплексов для повышения достоверности информации;

– разработаны алгоритмы повышения достоверности информации на основе применения ЭС, БД, БЗ, которые основаны на использовании логических и семантических связей элементов, характеристик и особенностей документов;

– рекомендовано использование прямых структурно-технологических, перекрестных взаимосвязей между элементами и отношений концептов ЭД, которые способствуют проектированию методов, существенно повышающих эффективность алгоритмов при восстановлении разрушенной или потерянной информации;

– смоделированы в структуре технологической схемы повышения достоверности информации ЭД механизмов структуризации, индексации и идентификации входных ЭД;

– рекомендована методика формирования БД, БЗ и реализации СУБД, правила контроля достоверности элементов ЭД в которых основаны на использовании операции реляционной алгебры.

Литература

[1] Об электронном документообороте. Закон Республики Узбекистан. 2004. №. 611-II.

[2] Бессонов С.В. Оптимизация электронного документооборота в корпоративных системах : дис. канд. экон. наук. 2000. С. 187.

[3] Келдыш Н.В. Методические основы автоматизированного решения задач ведомственного электронного документооборота // Науч. метод. сборник, 2005. №50. С. 110-117.

[4] *Гудов А. М.* Об одной модели оптимизации документопотоков, реализуемой при создании системы электронного документооборота. // сборник «Вычислительные технологии», 2006. №3. С. 53–65.

[5] *Абдуллаев Д. А., Амирсаидов У. Б.* Комплексная модель физического и канального уровней сети передачи данных // Вестник ТУИТ. Ташкент, 2007. №4. С. 19–43.

[6] *Жуманов И. И., Ахатов А. Р.* Алгоритм контроля качества текстов в системах электронного документооборота // Журнал «Вестник ТУИТ». — Ташкент, 2007.№2. С. 68–72.

[7] *Жуманов И. И., Ахатов А. Р.* Повышение надежности программного обеспечения проблемных задач в системах электронного документооборота // Материалы респ. науч. конф. «Современное состояние и пути развития информационных технологий», НТЦ «СИТ».. — Ташкент, 2006. С. 164–167.

[8] *Жуманов И. И., Ахатов А. Р.* Оценка эффективности программного комплекса контроля достоверности текстовой информации систем электронного документооборота // НТЖ «Химическая технология. Контроль и управление». — Ташкент, 2009.№2. С. 46–52.

[9] *Jumanov I. I., Axatov A. R.* Estimation of reliability for the system of mistakes dynamiic control at transfer and processing of the text information // Abstracts of Plenary and Invited Lectures of International School and Conference on Foliations, Dynamical Systems, Singularity Theory and Perverse Sheaves. — Samarkand, 2010.P. 80–85.

[10] *Гаврилова В. Т. А., Хорошевский В. В.* Базы знаний интеллектуальных систем // СПб: Питер. — 2000. С. 384.

[11] *Лукашевич Н. В.* Тезаурусы в задачах информационного поиска // М.: Изд-во Московского университета. — 2010. С. 512.

[12] *Jumanov I. I., Tursinxanov N. M., Axatov A. R.* Fuzzy Semantic Hypernet for Information Authenticity Controlling in Electronic Document Circulation Systems // -th International Conference on Application of Information and Communication Technologies, 12-14 october 2010, Section 2, IEEE. — Tashkent, 2010.P. 21–25.

[13] *Jumanov I. I., Axatov A. R.* Methods and algorithms of input information protection in electronic document processing systems // 2nd IEEE/IFIP International Conference ICI-2006, Uzbekistan, http: //ieeexplore.ieee.org/xpl/RecentCon.jsp?punumber=4055161. — Tashkent, 2006.

[14] *Jumanov I. I., Axatov A. R., Djumanov O. I.* An Effective Quality Control of Textual Information on the Basis of Statistical Redundancy in Distributed Mobile IT Systems and

e-Applications // 3-d International Conference in Central Asia on Internet, IEEE Catalog Number: 07EX1695C, ISBN: 1-4244-1007-X, Library of Congress: 2007920881. — Tashkent, 2007.

[15] *Жуманов И. И., Ахатов А. Р.* Оценки достоверности передачи изображений элементов текста в телекоммуникационных сетях // НТЖ «Химическая технология. Контроль и управление» -ТГТУ. — Ташкент, 2010. № 2. С. 23–30.

[16] *Жуманов И. И., Ахатов А. Р.* Оптимизация контроля передачи и обработки информации на базе технологии параллельных вычислений CUDA // НТЖ «Химическая технология. Контроль и управление» -ТГТУ. — Ташкент, 2009. № 5. С. 33–39.

[17] *Жуманов И. И., Ахатов А. Р.* ОАлгоритмы контроля достоверности изображений элементов текста в структуре пакетов передачи данных // НТЖ «Химическая технология. Контроль и управление» -ТГТУ. — Ташкент, 2010. № 3. С. 39–46.

[18] *Жуманов И. И., Ахатов А. Р.* Концептуальные принципы и методы контроля достоверности информации в структуре пакетов передачи данных на основе статистической избыточности // //«Илмий тадқиқотлар ахборотномаси» илмий-назарий, услубий журнал. — Самарканд, 2013. № 1(77). С. 39–49.

[19] *Jumanov I. I., Karshiev X. B.* Expanding the possibilities of instruments to improve the information reliability of electronic documents of industrial management SYSTEMS // Наука и мир, 2018. №3(55). С. 49–51.

[20] *Жуманов И. И., Каршиев Х. Б.* Методы обеспечения достоверности электронных документов на основе структурной избыточности и лексикологического синтеза // Tenth World Conference "Intelligent Systems for Industrial Automation", WCIS-2018. — Tashkent, 2018. P. 312–316.

[21] *Подлазова А. В.* Генетические алгоритмы в задачах плоского регулярного раскроя // Тр. II Междунар. конф. «Параллельные вычисления и задачи управления (PACO'2004)» / Ин-т пробл. упр, 2004. P. 284–316.

[22] *Курейчик В. М., Родзин С. И.* Эволюционные алгоритмы: генетическое программирование // Известия академии наук. Теория и системы управления, 2002. № 1. С. 127–137.

[23] *Куприянов М. С., Матвиенко Н. И.* Генетические алгоритмы и их реализации в системах реального времени // Информационные технологии, 2001. № 1. С. 17–21.

[24] *Липницкий А А.* Применение генетических алгоритмов к задаче о размещении прямоугольников // Кибернетика и системный анализ, 2002. № 6. С. 180–184.

Поступила в редакцию 05.04.2019

UDC 658.512.011

OPTIMIZATION OF INFORMATION ACCURACY BASED ON ELECTRONIC DOCUMENTS BASIS AND PECULIARITIES OF KNOWLEDGE BASE CONTROL RULES

Jumanov I. I., Karshiev H. B.

h-qarshiyev@samdu.uz;

Samarkand State University

In this article has been formulated problem of increasing the information reliability in electronic document management systems based on the use of algorithms based on performing typical functions of search, recognition, classification, generation, translation

of texts from one language to another, as well as monitoring and correcting errors of various multiplicity. The use of information redundancy of various nature, in particular, statistical, natural, structural, technological, semantic, has been substantiated and conceptual principles, methods, algorithms and software have been developed to increase information reliability. The control principle of the fidelity of elements, key concepts (words, phrases, terms) is proposed by comparing the entered document with the reference document - the original, as well as evaluating the information reliability based on the discrepancy coefficient. Algorithms have been developed that use logical, semantic, and structural-technological links between elements of document, moreover the tools have been obtained for using cross-relationships between individual records or groups of records as part of an information frame. Instruments of information control are designed based on selected elements of the knowledge base (KB) and the use of expert systems (ES). Algorithms for optimizing the allocation of ED with a variety of elements, attributes, concepts, fractal characteristics in databases (DB) and KB with the use of genetic search algorithms to select a document with the desired characteristics are proposed. The analysis results of proximity measure of document elements depending on the volume of the processed information and the probability of not detecting errors by the synthesized algorithm are obtained, and the effectiveness of the technology according to the complexity of information processing is investigated.

Keywords: electronic document, information accuracy, binary metric, relevance, search, recognition, document classification, program modules, performance indicators, generalized data processing algorithm.

Citation: Jumanov I. I., Karshiev H. B. 2019. Optimization of information accuracy based on electronic documents basis and peculiarities of knowledge base control rules. *Problems of Computational and Applied Mathematics.* 3(21): 57–74.