На правах рукописи

КУЗНЕЦОВ АНДРЕЙ ВИКТОРОВИЧ

АВТОМАТИЗАЦИЯ КОНТРОЛЯ ДОСТОВЕРНОСТИ ИНФОРМАЦИИ В ДОКУМЕНТАХ НА БУМАЖНЫХ НОСИТЕЛЯХ

05.13.06 – Автоматизация и управление технологическими процессами и производствами (промышленность)

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Орел 2012

Работа выполнена на кафедре «Информационные системы» в федеральном государственном бюджетном образовательном учреждении высшего профессионального образования «Государственный университет – учебно-научно-производственный комплекс».

Научный руководитель: доктор технических наук, профессор
Константинов Игорь Сергеевич

Официальные оппоненты: Поляков Александр Александрович
доктор технических наук, профессор,
МГУ им. М.В. Ломоносова, профессор кафедры математических методов в управлении

Архипов Олег Петрович
кандидат технических наук, старший научный сотрудник, ОФ ИПИ РАН, директор

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Белгородский государственный национальный исследовательский университет»

Защита состоится « 20 » марта 2012 г. в 16-00 часов на заседании диссертационного совета Д212.182.01 при ФГБОУ ВПО «Госуниверситет – УНПК» по адресу: 302020, РФ, г. Орел, Наугорское шоссе, д. 29, аудитория 212.

С диссертацией можно ознакомиться в библиотеке ФГБОУ ВПО «Госуниверситет – УНПК».

Автореферат разослан « 17 » февраля 2012 г.

Ученый секретарь
диссертационного совета Д 212.182.01
кандидат технических наук, доцент _____ В. Н. Волков

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Информационное обеспечение деятельности аппарата управления, его документирование, хранение и использование ранее созданных документов на предприятии реализуется посредством документооборота. При этом основным инструментом в документообороте является документ, и содержащаяся в нем информация.

Развитие электронных технологий послужило толчком к появлению электронного документооборота, однако документы на бумажных носителях (ДНБН) по-прежнему представляют большую ценность, а «электронные архивы» образуют дублирующую систему. И если современные способы защиты электронных документов (ЭД) близки к совершенству, то достоверность информации, содержащейся в ДНБН, находится на низком уровне по причине возросшего влияния человеческого фактора. Нарушение целостности информации в таких документах наиболее вероятно в процессе доставки оригинала текстового документа после прохождения им всех согласований, подписаний и утверждений ответственным лицам. Существующие на сегодняшний день методы полиграфической защиты ДНБН (Burrer F, Kezer K., Архипов О.П., Бородина Л.Н., Зыкова З.П., Богданова В.Н., Барсукова В.С., Иванова М.А.), в основном обеспечивают требуемую достоверность информации, и в большинстве случаев не оправдывают себя по причине дороговизны и узкой специализации реализующих их средств.

Целесообразным подходом в возникшей ситуации является использование для обеспечения контроля достоверности информации, содержащейся в ДНБН, традиционных средств офисной техники (ПЭВМ, лазерного принтера и планшетного сканера), а актуальным является разработка метода, методики и средств обеспечения контроля достоверности информации, содержащейся в ДНБН, реализующих установление схожести документа при его получении для ознакомления ответственным исполнителем с цифровой копией его оригинала, согласованного, подписанного и утвержденного ответственными лицами оригинала, единожды занесенного в электронную базу данных.

Таким образом, указанные обстоятельства и имеющиеся научные предпосылки обуславливают актуальность темы, объекта, предмета и цели диссертационного исследования.

Объект исследования – текстовая информация в документах предприятия на бумажных носителях.

Предмет исследования – методы, модели и алгоритмы оценки и обеспечения достоверности информации, содержащейся в документах на бумажных носителях, циркулирующих в системах документооборота предприятия.

Цель исследования – обеспечение достоверности информации, содержащейся в документах на бумажных носителях, и оперативности ее оценки.

Для достижения сформулированной цели были поставлены и решены ***следующие задачи.***

1. Анализ средств контроля информации содержащейся в документах предприятия.
2. Исследование критериев и методов оценки достоверности текстовой

информации, содержащейся в документах на бумажных носителях.

3. Разработка и исследование модели документа на бумажном носителе.

4. Разработка алгоритма оценки и создание методики автоматизированного контроля достоверности информации, содержащейся в документах на бумажных носителях, в системах документооборота предприятия.

5. Программная реализация прототипа системы обеспечения достоверности информации, содержащейся в документах на бумажных носителях.

Методы и средства исследования. В ходе исследования были использованы методы математической статистики, имитационного моделирования, теорий цифровой обработки сигналов, распознавания образов, вероятностей и случайных процессов. В разработке программного обеспечения использовалась технология применяемая в MATLAB в частности в приложении Image Toolbox.

Обоснованность и достоверность научных положений, основных выводов и результатов диссертации обеспечивается за счет тщательного анализа состояния исследований в данной области, и подтверждается корректностью предложенных модели, алгоритмов и методики, согласованностью результатов, полученных при исследовании алгоритма контроля достоверности бумажных документов, апробацией основных теоретических положений диссертации в печатных трудах и докладах на международных научных конференциях, а также в патентных предложениях.

Научная новизна диссертационного исследования заключается в том, что получены новые научные результаты:

1. *Структурная модель документа на бумажном носителе*, представляющая оцифрованный текстовый документ в виде иерархии морфологических признаков;
2. *Алгоритмы выделения структурных признаков* текстового документа на основе цифровой обработки изображения;
3. *Алгоритм оценки достоверности информации*, содержащейся в документах на бумажных носителях, без распознавания символов;
4. *Методика контроля достоверности информации*, содержащейся в документах на бумажных носителях, построенная на разработанных модели и алгоритмах.

Практическая ценность работы заключается в использовании теоретических результатов и разработанного программного модуля в автоматизированной системе документооборота предприятия (организации), по обеспечению достоверности получаемой информации на бумажных носителях.

Полученные теоретические результаты использованы:

1) в процессе обработки внутренних документов, а также договоров с внешними организациями в ЗАО «НАУЧПРИБОР» и ОАО «ОРЕЛАГРОПРОМСТРОЙ» (г. Орел);
2) в учебном процессе на кафедрах «Информационные системы» ГУ-УНПК, на кафедре «Радиотехника и электроника» академии ФСО России;
3) в разработке способа установления подлинности оригиналов бумажных документов (положительное решение по результатам формальной экс-

пертизы по заявке на изобретение № 2011131428 от 26.07.2011 г.);

4) в разработке программного средства морфологической обработки текстовых документов (свидетельство о государственной регистрации программы для ЭВМ № 2011619222 от 30.11 2011 г.);
5) в разработке системы считывания изображения (патент на полезную модель № 112790 от 20.01.2012).

Апробация и публикации. Отдельные результаты диссертационного исследования докладывались на: 3-й международной научно-практической конференции «Наука и бизнес: пути развития» (2011 г. Тамбов), XVI всероссийской научно-технической конференции в Рязанском Государственном радиотехническом университете (2011, г. Рязань), международной заочной научно-технической конференции «Современные тенденции в науке: новый взгляд» (2011, г. Тамбов), международной научно-практической интернет-конференции «Информационные технологии» (2011, г. Орел).

По материалам диссертации опубликовано 3 статьи в журналах из перечня ВАК, получено одно свидетельство о регистрации программы для ЭВМ, один патент на полезную модель и одной заявки на предполагаемое изобретение.

Положения, выносимые на защиту:

1. Структурная модель документа на бумажном носителе.
2. Алгоритмы выделения признаков.
3. Алгоритм оценки достоверности информации, содержащейся в документах на бумажных носителях.
4. Методика обеспечения достоверности информации, содержащейся в документах на бумажных носителях.

Структура и объем работы. Диссертационная работа изложена на 198 страницах и состоит из введения, четырех глав, заключения, списка литературы из 135 наименований и 4 приложений; содержит 9 таблиц и 37 рисунков.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность работы, сформулированы ее цель, задачи, научная новизна, практическая значимость и основные положения, выносимые на защиту.

Первая глава посвящена анализу организации документооборота на предприятии, существующих способов хранения и защиты документов, а также автоматизированных систем контроля и сопровождения документооборота.

Проведенный анализ показал, что электронный документооборот позволяет организованно подойти к решению проблем, связанных с обеспечением жизненного цикла документов, в частности, практически решенной задачей является задача обеспечения защищенности информации, содержащейся в ЭД, и физической сохранности ее носителей. Однако наряду с ним в значительных объемах существует оборот ДНБН, для которых применение средств защиты и определение достоверности содержащейся в них информации не всегда оправдано сточки зрения экономических и временных затрат.

Наиболее уязвимым местом в системе оборота ДНБН остается целостность информации (нарушаемая при несанкционированном уничтожении, добавлении лишних элементов и модификации данных), а существующие автоматизированные системы контроля и сопровождения документооборота в достаточной степени не решают вопросы контроля достоверности и соответствующих механизмов принятия решений.

Таким образом, совершенствование известных и разработка новых методов, методик и средств защиты документов, с возможностью определения достоверности содержащейся в них информации является актуальной проблемой, требующей глубоких исследований.

Во второй главе исследованы способы хранения оцифрованных документов в электронном виде, а также критерии и способы оценки достоверности информации, содержащейся в документах на бумажных носителях, сформулирована задача оценки достоверности такой информации.

Обеспечение достоверности информации, циркулирующей в системе бумажного документооборота, связана с необходимостью хранения оригиналов ДНБН. Анализ результатов сканирования текстовых документов, сохранения и обработки их в различных форматах показал, что для получения требуемого результата при оценке достоверности информации достаточным является использование формата JPEG с разрешением от 150 до 200 dpi.

Для оценки достоверности информации, содержащейся в ДНБН, при его получении для ознакомления исполнителем с хранимым оригиналом, согласованным, подписанным и утвержденным ответственными лицами, необходимо их сравнение по определенным критериям.

В работе для определения критериев оценки достоверности информации, содержащейся в текстовых документах, использованы структурные меры, учитывающие только дискретное строение данного информационного комплекса, в частности, количество содержащихся в нем информационных элементов. Предложено оценивать достоверность с использованием коэффициента схожести:

$$K_{\text{схож}} = \frac{N_{\Sigma\,\text{копия}} - N_{\Sigma\,\text{изм. копия}}}{N_{\Sigma\,\text{оригинал}}}, \qquad (1)$$

где $N_{\Sigma\,\text{оригинал}}$ – информационная емкость (общее число символов) текстового документа, занесенного в базу данных организации после прохождения им всех согласований, подписаний и утверждений ответственными лицами (оригинала); $N_{\Sigma\,\text{копия}}$ – информационная емкость текстового документа, полученного и отсканированного исполнителем (копия); $N_{\Sigma\,\text{изм. копия}}$ – число информационных элементов (символов) в копии, отличающихся от информационных элементов оригинала, и его асимптотическая оценка – вероятность схожести $P_{\text{схож}} = \lim_{N_{\Sigma\,\text{оригинал}} \to \infty} K_{\text{схож}}$.

Проанализированы существующие методы сравнения оцифрованных (отсканированных) документов, включающие в себя сравнение с эталоном, попиксельное сравнение, сравнение наложением и наложением со смещением. Указанные способы не обеспечивают достаточной точности оценки достоверности информации, содержащейся в текстовых документах, при приемлемой вычис-

лительной сложности алгоритмов сравнения, что ограничивает их применение в автоматизированных системах контроля и сопровождения документооборота.

Задача оценки достоверности информации сводится к сравнению копии и оригинала бумажного документа по выбранному критерию (1), и может трактоваться как одна из задач распознавания образов. Процесс распознавания в этом случае состоит в том, что на основании сопоставления апостериорной информации относительно каждого поступившего на вход системы объекта (отсканированного бумажного документа) с априорным описанием единственного класса, соответствующего оцифрованному оригиналу, принимается решение о принадлежности этого объекта к указанному классу. Формальная постановка задачи распознавания в рассматриваемом случае выглядит следующим образом: пусть задано множество объектов $\Omega = \{\omega_1,..., \omega_r\}$, представляющих собой цифровые изображения, полученные в результате многократного сканирования оригинала текстового документа и его копий, содержащих различные варианты частичной подделки (частичная или полная замена слов, строк, абзацев и страниц текста), пусть также определено множество возможных решений $L = \{l_1,..., l_k\}$, которые могут быть приняты системой, где решения $l \in L$ определяют степень отличия копии документа от его оригинала.

Для решения поставленной задачи разработана структурная модель текстового документа, построенная на основе:

1) исходного множества объектов $\Omega = \{\omega_1,..., \omega_r\}$;

2) множества возможных решений $L = \{l_1,..., l_k\}$;

3) априорного словаря признаков $x_a = \{x_1,..., x_N\}$;

4) меры близости объектов;

5) значений выигрышей, получаемых от принятия конкретных решений из множества $L = \{l_1,..., l_k\}$;

6) величины временных ресурсов T_o, ассигнованных на осуществление процедур выделения признаков.

В третьей главе разработаны алгоритмы оценки признаков априорного словаря системы оценки достоверности информации, содержащейся в документах на бумажных носителях, в качестве ее решения выбран метод математического моделирования, построен рабочий словарь детерминированных признаков и разработан алгоритм оценки достоверности информации, содержащейся в документах на бумажных носителях, а также проведено его исследование.

В качестве признаков априорного словаря формализованы следующие структурные компоненты текстового документа:

1) количество строк $N_{\text{строк}}$;

2) номера неполных строк $N_{\text{непол.строк}} = \lfloor n_{\text{непол.строк1}},..., n_{\text{непол.строк}i} \rfloor$, где $1 \le n_{\text{непол.строк}} \le N_{\text{строк}}$ – элемент вектора, соответствующий порядковому номеру неполной строки (как правило, первая и последняя строка абзаца);

3) количество слов в каждой строке $N_{\text{слов}} = \lfloor n_{\text{слов1}},..., n_{\text{слов}\,N_{\text{строк}}} \rfloor$, где $n_{\text{слов}\,i}$ – элемент вектора, равный числу слов в i-й строке;

4) расположение коротких слов $P_{\text{корот.слов}} = \left(p_{ij}\right)_{i=1,\,j=1}^{m,N_{строк}}$, где элемент вектора

$$P_{\text{корот. слов}} = \begin{cases} 1, \text{если } j \text{ - слово в } i \text{ - строке короткое} \\ 0, \text{в противном случае} \end{cases}$$

$i = 1...N_{\text{строк}}$, $j = 1..m$, m – максимальное число слов в строке анализируемого текста (строки с меньшим числом слов дополнялись справа нулями);

5) площадь слов $S = \left(s_{ij}\right)_{i=1,\,j=1}^{m,N_{строк}}$, где элемент матрицы s_{ij} – площадь (число пикселов области) j-го ($j = 1..m$) слова в i-й ($i = 1...N_{\text{строк}}$) строке (строки с числом слов меньшим m дополнялись справа нулями);

6) относительное расстояние между словами $L = \left(l_{ij}\right)_{i=1,\,j=1}^{m,N_{строк}}$, где элемент матрицы $l_{ij} = \sqrt{\left(x_{ij} - x_{11}\right)^2 + \left(y_{ij} - y_{11}\right)^2}$; x_{ij} и y_{ij} – горизонтальная и вертикальная координаты центра масс j-го ($j = 1..m$) слова в i-й ($i = 1...N_{\text{строк}}$) строке соответственно;

7) количество отверстий в буквах (по строкам) $N_{\text{отв}} = [n_{отв_1},...,n_{отв_{N_{строк}}}]$, где $n_{\text{отв. }i}$ – элемент вектора, равный числу отверстий в буквах слов i-й строки;

8) относительное расстояние между отверстиями в буквах $L_{отв} = \left(l_{отв_{ij}}\right)_{i=1,\,j=1}^{o,N_{строк}}$, где элемент матрицы $l_{\text{отв.}ij} = \sqrt{\left(x_{\text{отв.}ij} - x_{\text{отв.}11}\right)^2 + \left(y_{\text{отв.}ij} - y_{\text{отв.}11}\right)^2}$, $x_{\text{отв.}ij}$ и $y_{\text{отв.}ij}$ – горизонтальная и вертикальная координаты центра масс j-го ($j = 1..o$) отверстия в i-й ($i = 1...N_{\text{строк}}$) строке; o – максимальное число отверстий в строке анализируемого текста;

9) количество вертикальных линий в буквах $N_{\text{верт}} = [n_{верт_1},...,n_{верт_{N_{строк}}}]$, где $n_{\text{верт.}i}$ – элемент вектора, равный числу вертикальных линий в словах i-й строки;

10) относительное расстояние между вертикальными линиями в буквах $L_{верт} = \left(l_{верт_{ij}}\right)_{i=1,\,j=1}^{v,N_{строк}}$, где элемент матрицы $l_{\text{верт.}ij} = \sqrt{\left(x_{\text{верт.}ij} - x_{\text{верт.}11}\right)^2 + \left(y_{\text{верт.}ij} - y_{\text{верт.}11}\right)^2}$, $x_{\text{верт.}ij}$ и $y_{\text{верт.}ij}$ – горизонтальная и вертикальная координаты центра масс j-й ($j = 1..v$) вертикальной линии в i-й ($i = 1...N_{\text{строк}}$) строке; v – максимальное число вертикальных линий в строке анализируемого текста.

Для рассмотренных признаков в среде MATLAB реализованы соответствующие алгоритмы их оценивания, основанные на морфологической обработке оцифрованного текстового документа, в частности на операциях дилатация и эрозия. Для разработанных алгоритмов проверены их основные свойства и определена вычислительная сложность.

В качестве меры близости признаков оригинала и копии документа выбрана эвклидова метрика:

$$d^2(w, w_1) = \left(N_{\text{строк}}^{(p,k)} - N_{\text{строк}}^{(q,l)}\right)^2 + \sum_{j=1}^{t}\left(n_{\text{неполн.строк }j}^{(p,k)} - n_{\text{неполн.строк }j}^{(q,l)}\right)^2 +$$

$$+ \sum_{j=1}^{N_{\text{строк}}}\left(n_{\text{слов }j}^{(p,k)} - n_{\text{слов }j}^{(q,l)}\right)^2 + \sum_{j=1}^{N_{\text{строк}}}\sum_{i=1}^{m}\left(p_{\text{корот.слов }j,i}^{(p,k)} - p_{\text{корот.слов }j,i}^{(q,l)}\right)^2 + \sum_{j=1}^{N_{\text{строк}}}\sum_{i=1}^{m}\left(s_{\text{слов }j,i}^{(p,k)} - s_{\text{слов }j,i}^{(q,l)}\right)^2 +$$

$$+\sum_{j=1}^{N_{\text{строк}}}\sum_{i=1}^{m}\left(l_{\text{слов}\,j,i}^{(p,k)}-l_{\text{слов}\,j,i}^{(q,l)}\right)^2+\sum_{j=1}^{N_{\text{строк}}}\left(n_{\text{отв.}\,j}^{(p,k)}-n_{\text{отв.}\,j}^{(q,l)}\right)^2+\sum_{j=1}^{N_{\text{строк}}}\sum_{i=1}^{o}\left(l_{\text{отв.}\,j,i}^{(p,k)}-l_{\text{отв.}\,j,i}^{(q,l)}\right)^2+$$
$$+\sum_{j=1}^{N_{\text{строк}}}\left(n_{\text{верт.}\,j}^{(p,k)}-n_{\text{верт.}\,j}^{(q,l)}\right)^2+\sum_{j=1}^{N_{\text{строк}}}\sum_{i=1}^{o}\left(l_{\text{верт.}\,j,i}^{(p,k)}-l_{\text{верт.}\,j,i}^{(q,l)}\right)^2=d_1^2\left(w_{pk},w_{ql}\right)+d_2^2\left(w_{pk},w_{ql}\right)+$$
$$+d_3^2\left(w_{pk},w_{ql}\right)+d_4^2\left(w_{pk},w_{ql}\right)+d_5^2\left(w_{pk},w_{ql}\right)+d_6^2\left(w_{pk},w_{ql}\right)+d_7^2\left(w_{pk},w_{ql}\right)+$$
$$+d_8^2\left(w_{pk},w_{ql}\right)+d_9^2\left(w_{pk},w_{ql}\right)+d_{10}^2\left(w_{pk},w_{ql}\right), \qquad (2)$$

где $d_j^2(w,w_1)$, $j=1...10$ – эвклидова мера близости параметров $N_{\text{строк}}$, $N_{\text{неполн.строк}}$, $N_{\text{слов}}$, $P_{\text{корот.слов}}$, $S_{\text{слов}}$, $L_{\text{слов}}$, $N_{\text{отв}}$, $L_{\text{отв}}$, $N_{\text{верт}}$, $L_{\text{верт}}$ оригинала и копии текстового документа.

Учитывая различный вклад отдельных параметров в формирование общей меры близости (2) и необходимость сокращения априорного словаря, совокупность признаков объектов, используемых в рабочем словаре, описана N-мерным вектором $\Lambda=\{l_1,l_2,...,l_N\}$, компоненты которого определяют вес соответствующего признака. С учетом Λ квадрат расстояния между объектами (2) составил:

$$d^2\left(w_{pk},w_{ql}\right)=l_1d_1^2\left(w_{pk},w_{ql}\right)+l_2d_2^2\left(w_{pk},w_{ql}\right)+l_3d_3^2\left(w_{pk},w_{ql}\right)+l_4d_4^2\left(w_{pk},w_{ql}\right)+l_5d_5^2\left(w_{pk},w_{ql}\right)+$$
$$+l_6d_6^2\left(w_{pk},w_{ql}\right)+l_7d_7^2\left(w_{pk},w_{ql}\right)+l_8d_8^2\left(w_{pk},w_{ql}\right)+l_9d_9^2\left(w_{pk},w_{ql}\right)+l_{10}d_{10}^2\left(w_{pk},w_{ql}\right). \qquad (3)$$

Для определения значений выигрышей, получаемых от принятия конкретных решений из множества $L=\{l_1,...,l_k\}$ и определения коэффициентов l_j, $j=1,2,...,10$, в работе использовался множественный регрессионный анализ.

Важным условием его применения является независимость и нормальность распределения независимых величин. Для оценки указанных свойств производилось 100-кратное сканирование оригинала текстового документа с разрешением 150 dpi и определение параметров априорного словаря. Анализ полученных результатов показал, нулевую дисперсию параметров $d_j^2(w,w_1)$ для $j=1...3$. Анализ выбросов, присутствовавших на гистограммах распределения оставшихся параметров, показал, что они свойственны изображениям, имеющим значительный поворот $(\geq 2^{\circ})$ относительно оригинала.

Анализ исследуемого множества изображений, исключая указанные, позволил аппроксимировать распределения мер близости (рис. 1) нормальным законом. Подбор распределений осуществлен с помощью средства *Distribution-FittingTool* (dfittool) пакета *Statistics* программы технических расчетов MATLAB. Для проверки гипотезы о нормальности распределения мер близости параметров априорного словаря признаков использовался критерий согласия Колмогорова-Смирнова. При этом для каждого распределения нулевая гипотеза состояла в том, что распределение генеральной совокупности не противоречит стандартному нормальному закону, а альтернативная – в том, что распределение генеральной совокупности противоречит стандартному нормальному закону. Для всех параметров на критическом уровне значимости $p_{кр}=0{,}05$ подтвердилась нулевая гипотеза.

В качестве зависимого параметра регрессионной модели выбран коэффициент $K_{\text{схож}}$ (1).

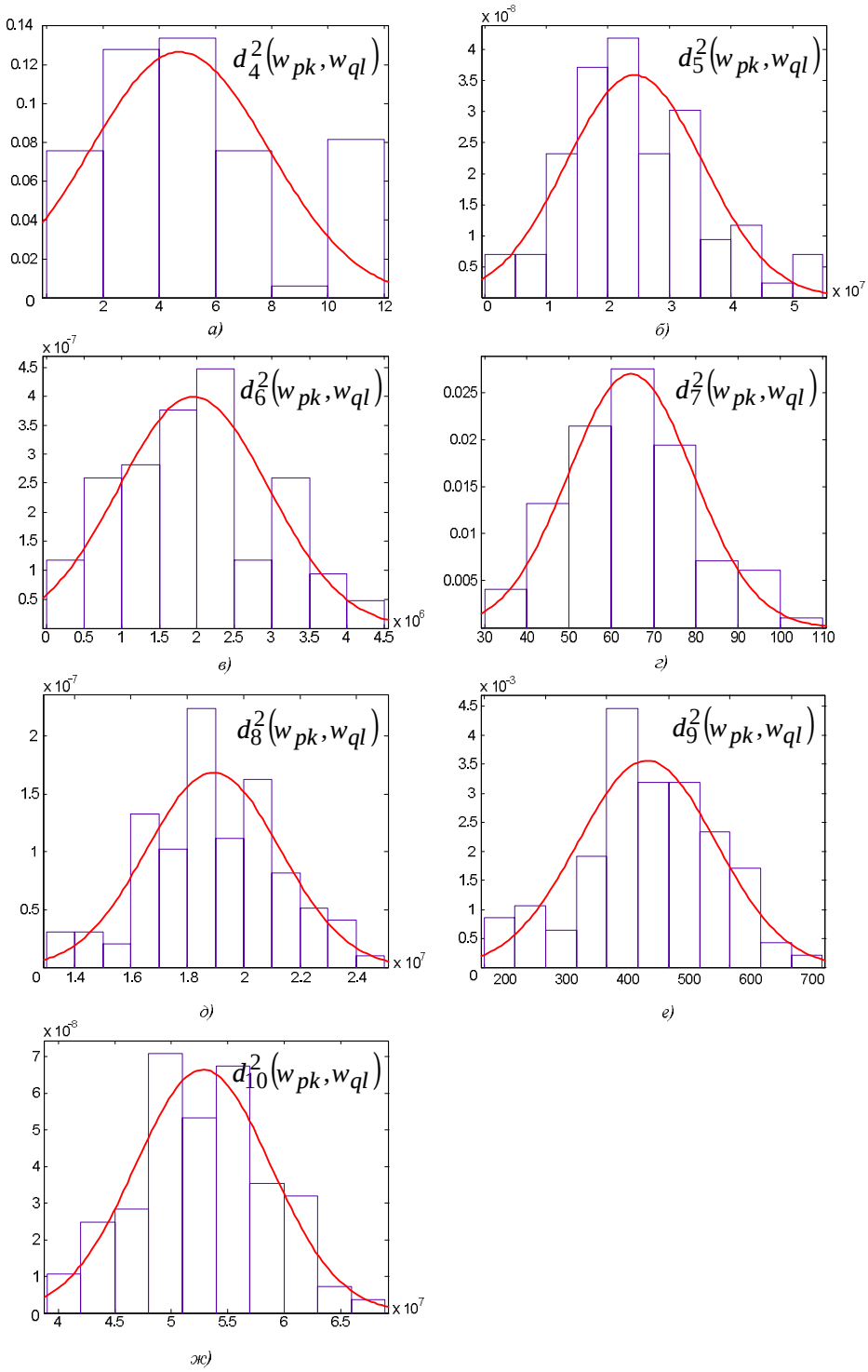

Рисунок 1 – Аппроксимация распределений значений мер близости параметров

В указанном случае зависимость выбранных критерия достоверности (1) и меры близости текстовых документов (3) определена как

$$K_{\text{схож}} = l_1 d_1^2(w, w_{11}) + l_2 d_2^2(w, w_{11}) + l_3 d_3^2(w, w_{11}) + l_4 d_4^2(w, w_{11}) + l_5 d_5^2(w, w_{11}) + l_6 d_6^2(w, w_{11}) + l_7 d_7^2(w, w_{11}) + l_8 d_8^2(w, w_{11}) + l_9 d_9^2(w, w_{11}) + l_{10} d_{10}^2(w, w_{11}) \quad (4)$$

Для определения коэффициентов регрессии (4) исследованию подвергались 2200 оцифрованных изображений, оригинала документа и его 30 копий, содержащих подделки.

В результате множественный регрессионный анализ позволил на уровне значимости $p = 0{,}0000001$ (по критерию Фишера) получить коэффициент множественной регрессии $R = 0{,}9868$ и следующие значения коэффициентов регрессии (табл. 1).

Таблица 1

λ_j	Вычисленные значения	Ошибки оценивания	Значения статистического критерия Стьюдента	Значения уровней значимости по критерию Стьюдента
λ_0	1,024285834707	0,018663847892	54,8807	0
λ_1	0,006560966142	0,003331107147	1,9696	0,0492
λ_2	0,000349635099	0,000024681726	14,1657	0
λ_3	0,000864616833	0,000125530532	6,8877	0
λ_4	–0,006807158182	0,001410106790	–4,8274	0
λ_5	–0,000000000541	0,000000000210	–2,5833	0,0099
λ_6	0,000000005455	0,000000000649	8,4030	0
λ_7	–0,000582593170	0,000021761645	–26,7716	0
λ_8	0,000000000953	0,000000000369	2,5837	0,0099
λ_9	0,000019461904	0,000006226881	3,1255	0,0018
λ_{10}	0,000000000385	0,000000000235	1,6397	0,1014

Полученные результаты позволяли признать значимыми признаки l_j для $j = 1...9$ (исключить признак $L_{\text{верт}}$) и, используя полученные результаты, представить структурную модель текстового документа в следующем виде:

$$K_{\text{схож}} = l_0 + \sum_{j=1}^{9} l_j d_j^2(w, w_{11}) \qquad (5)$$

Представленная модель (5) в работе проверена на адекватность и точность.

Полученная модель и алгоритмы определения структурных признаков позволили представить алгоритм оценки достоверности информации, содержащейся в текстовом документе следующим образом (рис. 2).

Для оценки точности разработанного алгоритма по результатам статистических испытаний использовалась средняя абсолютная ошибка

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

характеризующая точность алгоритма. Полученные результаты (для 3 оригиналов и 12 документов, содержащих частичную подделку) свидетельствуют о высокой точности ($MPAE \le 5\,\%$) разработанного алгоритма. Однако разброс оцениваемых параметров для копий документа, не содержащих подделки, указал на необходимость представлять множество возможных решений в виде:

$$L = \begin{cases} l_1 \Rightarrow \{K_{\text{схож}} \ge 0{,}97\}, \\ l_2 \Rightarrow \{K_{\text{схож}} < 0{,}97\}, \end{cases}$$

где решение l_1 указывает на соответствие копии документа его оригиналу с погрешностью метода оценивания 3 %, а l_2 – на любые случаи подделки, требующие принятия дополнительных мер по обеспечению достоверности.

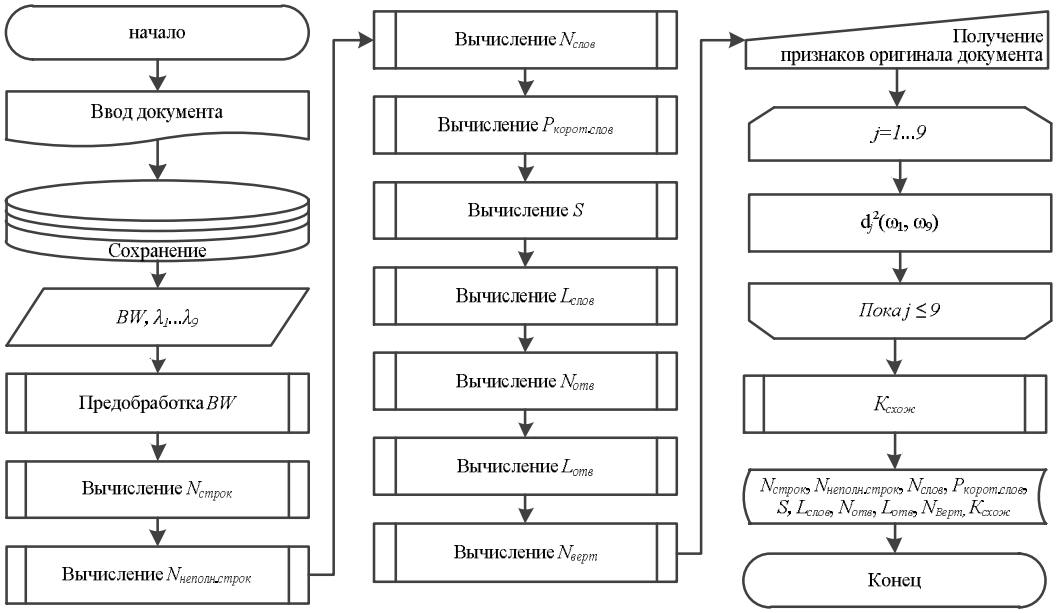

Рисунок 2 – Алгоритм оценки достоверности информации, содержащейся в ДНБН

Для сравнения полученных результатов с существующими проведен анализ совокупности программ ABBYY FineReader и Microsoft WORD, являющихся в настоящее время единственным средством сотрудника организации для сравнения документов. Полученные результаты (табл. 2) показали превосходство разработанного алгоритма по критерию времени оценивания достоверности (оперативности) и позволили установить требования к величине временных ресурсов в виде $T_0 = 1$ мин безотносительно элементной базы.

Таблица 2 – Оперативность оценки достоверности информации в ДНБН на ПЭВМ Р-4/1.8ГГц

Условный шифр документа	Microsoft WORD			Разработанный алгоритм	
	Время распознавания, с	Время сравнения, с	Общее время, с	Время обработки, с	Общее время, с
К.1	20,5	2,5	48	19,12	44,12
П.1.1	22,5	2,5	50	20,37	45,37
П.1.2	18,5	2,5	46	20,14	45,14
П.1.3	20,5	2,5	48	19,83	44,83
К.2	14,5	2,5	42	12,94	37,94
П.2.1	15,5	2,5	43	12,53	37,53
П.2.2	16	2,5	43,5	12,78	37,78
П.2.3	16,5	2,5	44	12,48	37,48
К.3	16	2,5	43,5	13,78	38,78
П.3.1	16	2,5	43,5	13,91	38,91
П.3.2	15,5	2,5	43	13,78	38,78
П.3.3	16	2,5	43,5	13,85	38,85
Среднее время:			**44,83**		**40,46**

Примечания: время сканирования одной страницы 25 с; время сохранения и распознавания одной страницы (ПЭВМ Core 2 Duo 3,3 ГГц / 4Гб) 1 с; время сохранения и распознавания одной страницы (ПЭВМ Pentium 4/1.8 ГГц/512Mb) 2,5 с.

Кроме того, следует учесть, что определения схожести документов с помощью WORD необходимо ручное вмешательство, а также для качественного распознавания текста с помощью FineReader необходимо разрешение сканирования 300dpi, что увеличивает общее время оценки.

Четвертая глава посвящена формированию функциональной схемы, состава системы, разработке программного комплекса ПДИБД и методики обеспечения достоверности информации, содержащейся в документах на бумажных носителях.

Для повышения достоверности информации, содержащейся в документах на бумажных носителях, и реализации разработанного алгоритма предложен следующий вариант функциональной схемы системы обеспечения достоверности (рис. 3).

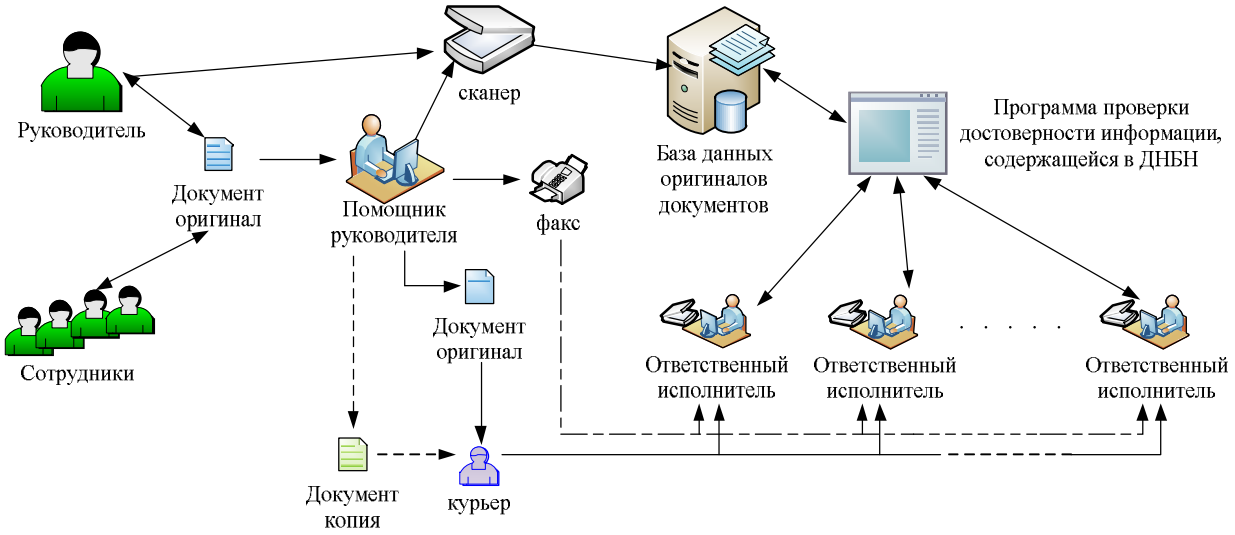

Рисунок – 3 Функциональная схема системы обеспечения достоверности информации, содержащейся в бумажных документах

Аппаратная часть данной системы реализована традиционными средствами офисной техники, локальной вычислительной сетью и защищенной базой данных, программная часть реализована в виде двухуровневой модели с разграничением прав доступа. Так, для помощника руководителя (руководителя) реализована возможность занесения оцифрованного документа и его признаков в защищенную базу данных организации (функции первого уровня), а также возможность оценки достоверности информации, содержащейся в текстовом документе (функции второго уровня). Ответственным исполнителям доступны функции второго уровня.

Результаты исследования показали на соответствие разработанного программного комплекса необходимым требованиям нормального функционирования.

Основываясь на разработанной функциональной схеме системы (рис. 3), предложена методика обеспечения достоверности информации, содержащейся в бумажных документах, включающая в себя следующие этапы.

1. Формирование образа оригинала. Оригинал текстового документа после прохождения им всех согласований, подписаний и утверждений ответственными лицами регистрируется помощником руководителя (руководителем) и заносится в защищенную базу данных организации (функции первого уровня).

2. *Контроль достоверности.* Получение документа ответственным исполнителем и оценка его достоверности с использованием разработанного программного комплекса (функции второго уровня).

Апробация разработанной методики показала повышение достоверности информации, содержащейся в ДНБН, и оперативности ее оценки.

В заключении сформулированы основные результаты работы.

В приложениях представлены результаты проведенных экспериментов.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ:

В диссертационной работе решена актуальная научно-техническая задача разработки методики обеспечения достоверности информации, содержащейся в документах на бумажных носителях, в системах документооборота предприятия.

В рамках проведенных исследований получены следующие ***основные результаты***:

1. Проведенный анализ существующих автоматизированных систем контроля и сопровождения документооборота показал, что сейчас в полной мере не решены вопросы защиты информации в документах на бумажных носителях и контроля ее достоверности, а существующие средства и методы защиты документов обеспечивают в основном целостность самого документа, но не информации, содержащейся в нем, что соответственно является слабым звеном в данной ситуации, т.к. возможна сознательная или случайная подделка документов, представленных в бумажном виде.

2. Предложен метод решения задачи построения системы оценки достоверности информации, содержащейся в документах на бумажных носителях.

3. Построен априорный словарь детерминированных структурных признаков, полученных в результате морфологической обработки цифровых изображений текстовых документов посредством разработанных оригинальных алгоритмов.

4. Учитывая статистическую взаимосвязь между параметрами априорного словаря на основе метода множественной регрессии разработан рабочий словарь признаков, учитывающий вклад частных мер близости признаков оригинала документа и его копий.

5. Разработана структурная модель текстового документа, позволяющая представить документ в виде иерархии морфологических признаков.

6. На основе представленной модели текстового документа предложен вариант алгоритма оценки достоверности информации, содержащейся в документах на бумажных носителях, в котором в качестве исходных данных используются цифровое изображение копии документа, получаемое в результате его сканирования, коэффициенты регрессии, и признаки оригинального документа.

7. На основе произведенных экспериментов получены результаты, свидетельствующие о целесообразности применения разработанного алгоритма с точки зрения точности получаемых результатов, времени выполнения, стоимости и эргономичности представления результатов.

8. По результатам статистических испытаний определено требование к величине временных ресурсов, ассигнованных на осуществление процедур выделения признаков.

9. Предложен вариант функциональной схемы системы обеспечения достоверности информации, содержащейся в документах на бумажных носителях, позволяющей повысить эффективность управленческого воздействия в системе функционирования документооборота в результате автоматизированного контроля.

10. Сформированы минимальные требования к аппаратной части ЭВМ для нормального функционирования системы.

11. Реализован прототип системы обеспечения достоверности информации содержащейся в документах на бумажных носителях в системе документооборота функциями двух уровней: руководителем (помощником руководителя) и ответственными исполнителями.

12. Представлена методика контроля достоверности информации, содержащейся в документах на бумажных носителях, гарантирующая достоверность от 97% , и обеспечивающая экономию времени на операциях сравнения документов и избавляющая сотрудников от ведения визуального контроля и перезапуска программ.

13. Результаты работы внедрены на предприятиях ЗАО «НАУЧПРИБОР» и ОАО «ОРЕЛАГРОПРОМСТРОЙ», а также в учебном процессе на кафедре «Информационные системы» Госуниверситета - УНПК, на кафедре «Радиотехника и электроника» академии ФСО России, опубликованы в 10 печатных трудах, докладах на конференциях и патентных предложениях.

Список работ, опубликованных по теме диссертации в изданиях, рекомендованных ВАК РФ

1. **Кузнецов, А. В.** Проблемы достоверности документов [Текст] / А. В. Кузнецов // Известия ОрелГТУ. Информационные системы и технологии. – Орел: ОрелГТУ, 2009. – № 1/51(562). – С. 51-57.

2. **Кузнецов, А. В.** Организация сопровождения жизненного цикла документов [Текст] / А. В. Кузнецов // Информационные системы и технологии – Орел: Госуниверситет - УНПК, 2011. – № 1/(63). – С. 68 – 72.

3. **Кузнецов, А. В.** Регрессионная модель разности структурных признаков текстовых документов [Текст] / А. В. Кузнецов, И. С. Константинов, О. О. Басов // Информационные системы и технологии. – Орел: Госуниверситет - УНПК, 2012. – № 1(69). – С. 114 – 123. *(Личное участие 50%)*

Список работ, опубликованных по теме диссертации в материалах конференций

4. **Кузнецов, А. В.** Способ определения схожести содержательной части документов. [Текст] / А. В. Кузнецов // 3-я международная научно-практическая конференция «Наука и бизнес: пути развития». Труды конференции. – Тамбов: Изд-во ТАМБОВПРИНТ, 2011. – 62 с. - С.50-52.

5. **Кузнецов, А. В.** Метод установления схожести содержательной части бумажного документа с цифровой копией его оригинала. [Текст] / А. В. Кузнецов, О. О. Басов // «Информационные системы и технологии». Материалы

международной научно-технической интернет конференции. г. Орел, апрель-май 2011. В 3 т. Т. 3 – Орел: ФГОУ ВПО «Госуниверситет-УНПК», 2011. – Т.3. – С. 67-71. *(Личное участие 50%)*

6. **Кузнецов, А. В**. Структурная модель текстового документа [Текст] / А.В. Кузнецов, О. О. Басов, И. В. Блинов // «Новые информационные технологии в научных исследованиях». Материалы XVI всероссийской научно-технической конференции студентов, молодых ученых и специалистов. г. Рязань, 2011. – С. 286-287. *(Личное участие 60%)*

7. **Кузнецов, А. В.** Методика обеспечения достоверности бумажных документов в системах документооборота. [Текст] / А. В. Кузнецов, О.О. Басов, И. В. Блинов // «Современные тенденции в науке: новый взгляд». Материалы международной заочной научно-технической конференции. г. Тамбов, 2011. – С. 72-74. *(Личное участие 50%)*

8. **Кузнецов, А. В.** Свидетельство об официальной регистрации программы для ЭВМ № 2011619222 «Морфологическая обработка текстовых документов» / А.В. Кузнецов, О.О. Басов. – Федеральная служба по интеллектуальной собственности, патентам и товарным знакам: Реестр программ для ЭВМ. – 30.11.2011. *(Личное участие 50%)*

9. **Кузнецов, А. В.** Патент на полезную модель № 112790 «Система считывания изображения» / А. В. Кузнецов, О. О. Басов, А. И. Офицеров, И. Ю. Баранов. – Федеральная служба по интеллектуальной собственности, патентам и товарным знакам: Государственный реестр полезных моделей Российской Федерации. – 20.01.2012. *(Личное участие 40%)*

10. **Кузнецов, А. В.** Положительное решение по результатам формальной экспертизы по заявке на изобретение № 2011131428 «Способ установления подлинности оригиналов бумажных документов» / А. В. Кузнецов, О. О. Басов, А. И. Офицеров. – Федеральная служба по интеллектуальной собственности, патентам и товарным знакам: Государственный реестр полезных моделей Российской Федерации. – 18.11.2011. *(Личное участие 50%)*

ЛР ИД № 00670 от 05.01.2000 г.
Подписано к печати « 14 » февраля 2012 г.
Усл. печ. л.1,00 Тираж 100 экз.
Заказ № 146.

Полиграфический отдел ФГБОУ ВПО «Госуниверситет – УНПК»
302005, г. Орел, ул. Московская, 65