ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
INFORMATICS AND INFORMATION TECHNOLOGIES

УДК 519.681.5

ПОВЫШЕНИЕ ДОСТОВЕРНОСТИ ИНФОРМАЦИИ НА ОСНОВЕ ЛОГИЧЕСКОЙ И СЕМАНТИЧЕСКОЙ ИЗБЫТОЧНОСТИ, СВЯЗЕЙ ЭЛЕМЕНТОВ И ОТНОШЕНИЙ КЛЮЧЕВЫХ КОНЦЕПТОВ ДОКУМЕНТА

Жуманов И.И., Каршиев Х.Б.

Разработаны методы повышения достоверности информации, основанные на использовании статистической, естественной, структурно-технологической, семантической и информационной избыточности. Предложены методики определения объема избыточности, требуемой при повышении достоверности информации за счет статистических, логических, семантических связей элементов и отношений концептов, а также расчета метрического расстояния между введенным и эталонным документами. Разработан метод на основе концептуального принципа контроля заголовков протоколов передачи данных в телекоммуникационных сетях с механизмами формирования эталона и проверки достоверности информации, фиксирования информационного поля пакета, выбора правило контроля. Предложен метод и алгоритм повышения достоверности информации на основе n-граммной модели узбекского языка с применением механизма кластеризации. Реализована схема формирования и применения частотного словаря словоформ, определения вероятности принадлежности проверяемого слова к классу достоверных по логарифмической функции. Разработан программный комплекс повышения достоверности информации на языке C#.NET, включающий алгоритмы повышения достоверности информации.

Ключевые слова: electronic document, information accuracy, set of standards, information redundancy, interconnectedness of elements, relation of concepts, clustering, n-gram model, software package.

Стастистик, табиий, таркибий-технологик, семантик ортиқчаликлардан фойдаланиш асосида, ахборот ишончлигини оширувчи усуллар ишлаб чиқилган. Элементлар орасида мавжуд стастистик, мантиқий, семантик алоқалар ва калит концептлар муносабатлари бўйича, ахборот ишончлигини оширишда талаб этилган ортиқчалик ҳажмини аниқловчи, ҳамда кириш ва эталон ҳужжатлари орасидаги метрик масофани ҳисобловчи услубиятлар таклиф этилган. Телекомуникация тармоқларида ахборотни узатиш протоколлари сарлавҳаларини назорат қилиш концептуал принципи,

эталонни шакллантириш ва ахборот ишончлигини текшириш, пакетнинг ахборот майдонини белгилаш, назорат қилиш қоидаларини танловчи механизмларга асосланган усул ишлаб чиқилган. Кластеризация механизмини қўллашга таянган, ўзбек тили n-граммли модели асосида, ахборот ишончлигини ошириш усул ва алгоритми таклиф этилган. Сўз шакли частотали луғатини шакллантириш ва қўллаш, логарифмик функция бўйича назорат қилинувчи сўзнинг ишончлилик синфига тегишлилигини аниқловчи схема жорийлаштирилган. C#.NET тили асосида, ҳамда информацияни назорат қилувчи алгоритмлардан таркиб топган ахборот ишончлигини оширувчи дастурий мажмуа яратилган.

***Таянч иборалар*:** электрон ҳужжат, информация ишончлилиги, эталон – жамлама, информация ортиқчалиги, элементлар боғликлиги, концептлар муносабати, кластерлаш, n-граммли модел, дастурий мажмуа.

In this article have been developed methods for increasing the reliability of information based on the use of statistical, natural, structural-technological, semantic information redundancy. Moreover, methods have been proposed for determining the amount of redundancy required for increasing the information reliability through statistical, logical, semantic relationships of elements and concept relationships, as well as calculating the metric distance between the input and the reference documents. A method has been developed based on the conceptual principle of monitoring the data transfer header protocols in telecommunication networks with mechanisms for generating a template and verifying the accuracy of information, fixing an information field of a packet, choosing a control rule. A method and algorithm for improving the information reliability based on the n - gram model of the Uzbek language using a clustering mechanism are proposed. A scheme has been implemented for the formation and application of the frequency dictionary of word forms, for determining the probability of the word being checked to belong to the class of reliable logarithmic functions. A software package for improving the reliability of information in the C#.NET has been developed, the software includes algorithms for increasing the information reliability.

***Keywords*:** electronic document, reliability of information, set-standard, information redundancy, interconnectedness of elements, relation of concepts, clustering, n-gram model, software package.

I. ВВЕДЕНИЕ

Актуальность темы. В системах электронного документооборота (СЭД), информационно - ресурсных фондов мониторинга производственно-технологических процессов наблюдается постоянный рост передаваемых документов, методы обработки которых связаны совмещением возможностей

типовых инструментов поиска, распознавания, классификации, повышения достоверности, сохранности, целостности информации [1,2]. Электронные документы (ЭД) представляются пользователю по локальным или глобальным сетям СЭД, в которых производятся поиск, отбор нужного документа по ключевым концептам, словам и рубрикаторам из огромного списка документов в базе данных (БД).

Традиционные подходы к повышению достоверности информации определяются широким классом кодовых и аппаратурных методов, основывающихся на использовании искусственной избыточности, которые образуются дополнительно включаемыми контрольными символами в кодовые комбинации передаваемых сообщений [3,4]. Применяемые в автоматизированных системах управления программные методы для повышения достоверности информации основаны на образовании в строке либо в столбце передаваемого к обработке документа дополнительных контрольных разрядов путем посимвольного, цифрового, модульного суммирования значений элементов концептов и слов.

Несмотря на большие преимущества по обеспечению достоверности информации эти методы считаются много затратными. В связи с этим, современные технологии обработки информации на предприятиях малого и среднего масштаба нуждаются в простых, устойчивых к ошибкам методах повышения достоверности информации ЭД [2].

В [5,6] исследован и разработан широкий спектр разнообразных методов повышения достоверности информации, которые основаны на использовании статистической, естественной, структурно-технологической, семантической информационной избыточности. При этом, инструменты обработки информации, усовершенствованные на этих концептуальных принципах, отвечают всем предъявляемым требованиям и решают задач с намного меньшими вычислительными затратами.

С понятием информационной избыточности связаны получение энтропийных характеристики документа, оценок количества информации, рационального объема избыточности, которые используются при выборе и проектирование методов повышения достоверности информации на основе избыточности различной природы.

В [4] решены задачи оценки упрощенной условной энтропии информации документов делопроизводства, основанные на формулах Д. Хартли и К.Шеннона.

В настоящей работе предложен подход, направленный на разработку методики определения рационального объема информационной избыточности в документах, методов обработки изображений на основе использования статистических, логических, семантических связей элементов и отношений концептов, а также фрактальных специфических характеристик документов, которые для повышения достоверности информации

представляются в формализованном виде, сравниваются с эталонным аналогом, определяется сходство введенного документа с эталонным на основе геометрической близости их точек [7].

II. ОСНОВНАЯ ЧАСТЬ

Концептуальные принципы использования информационной избыточности. Условная энтропия одного элемента концепта S_i документа определяется средней энтропией по алфавиту элементарных сообщений при условии, что известен предыдущий элемент $S_k^{/}$:

$$H(S_i/S_k^{/}) = -\sum_{k=1}^{2^m} P(S_k^{/}) \sum_{i=1}^{2^m} P(S_i/S_k^{/}) \log P(S_i/S_k^{/}). \qquad (1)$$

Для вычисления энтропии информации (1) документа необходимо определить многомерные вероятности вида $P(S_i/S_k^{/})$. В работе [7] дана методика оценки верхней границы условной энтропии, которая учитывает как неравномерность распределения, так и корреляцию между элементами проверяемого концепта документа. Исходя из этого, запишем оценку среднего количества информации о элементе концепта в виде:

$$J(S_i/S_k^{/}) = H(S_i) - H(S_i/S_k^{/}) = -\sum_{i=1}^{2^m} P(S_i) \log P(S_i) + \sum_{k=1}^{2^m} P(S_k^{/}) \sum_{i=1}^{2^m} P(S_i/S_k^{/}) \log P(S_i/S_k^{/})$$

Количество информации, требуемое для алгоритмов повышения достоверности информации в структуре пакета передачи данных (ППП) задаётся в виде:

$$J(S_i/S_k^{/}) = m + \frac{Q(S)}{1-P_0(S)} \log Q(S) - \frac{Q(S)}{1-P_0(S)} \log[1-P_0(S)] + \frac{P_H(S)}{1-P_0(S)} \times$$

$$\times \log P_H(S) - \frac{P_H(S)}{1-P_0(S)} \log[1-P_0(S)] - \frac{P_H(S)}{1-P_0(S)} \log(2^m - 1) = m - \frac{P_H(S)}{1-P_0(S)} \times$$

$$\times \log(2^m - 1) - \log[1-P_0(S)] + \frac{Q(S)}{1-P_0(S)} \log Q(S) + \frac{P_H(S)}{1-P_0(S)} \log P_H(S),$$

где $P_H(S) = 0$, $J(S_i/S_k^{/}) = m$;

$Q(S) = 1 - P_0(S)$ при другом граничном условии имеем $Q(S) = 0$;

$P_H(S) = 1 - P_0(S)$, при $J(S_i/S_k^{/}) = m - \log_2(2^m - 1)$;

$J(S_i/S_k^{/})$ - определяет тот минимум дополнительной информации, который необходим для повышения достоверности информации.

Методика определения требуемого объема избыточности. Будем считать, что элементы документа кодируются кодом *(n,k)*, где *n* – длина информационной части кода, а *k* – число контрольных символов.

Причем, чем больше число проверочных символов, тем лучше ошибка обнаруживающая способность кода. Для простоты рассмотрим код, когда $n=7$, а $k=3$.

Вероятность правильной проверки информации определяется, как:

$$Q(S)=(1-P)^m, \qquad m=10,$$

где P- средняя вероятность ошибок элемента в концепте документа.

Вероятность обнаружения ошибок запишется в виде:

$$P_0(S)=(1-P)^7+7P^4(1-P)^3.$$

Вероятность неправильной проверки информации определяется в виде

$$P_H(S)=(2^m-1)P^4(1-P)^3.$$

Эффективность вводимого алгоритма зависит от требуемого количество информации, определяемого в виде:

$$J(S_i/S_k^{/})=m\frac{(2^m-1)P^4(1-P)^3}{[1-(1-P)^7+7P^4(1-P)^3}\log(2^m-1)-\log[1-(1-P)^7+7P^4(1-P)^3+$$

$$+\frac{(1-P)^m}{[1-(1-P)^7+7P^4(1-P)^3}\log(1-P)^m+\log(2^m-1)-\log[1-(1-P)^7+7P^4(1-P)^3+$$

$$+\frac{(1-P)^m}{[1-(1-P)^7+7P^4(1-P)^3}\log(1-P)^m+\frac{(2^m-1)R^4(1-P)^m}{[1-(1-P)^7+7P^4(1-P)^3}\log(2^m-1)P^4(1-P)^3.$$

Учитывая, что средняя вероятность ошибок этапов ввода, передачи, хранения и обработки информации $P<<1$, выражение требуемого количество информации задаётся в виде

$$J(S_0/S_k^{/})=m-(2^m-1)\log(2^m-1)-\log 7P^4(1-P)+(2^m-1)\log[(2^m-1)P^4].$$

Получены оценки условной энтропии для концепта документа, задаваемого различным числом элементов.

Определен требуемый объем избыточности для методов повышения достоверности информации, основанных на использовании избыточностей различной природы в виде следующей формулы:

$$R(S_i)=1-\frac{1}{(1-P)}\log\frac{(2^m-1)}{nP},$$

которая позволяет оценить также вероятность необнаруженных ошибок при различных механизмах использования избыточности в допустимых пределах.

В табл.1 приведены экспериментальные результаты оценки избыточности информации элементов трех концептов, принадлежащих трем группам разнородных документов. Исследованию подвергались 100 документов, относящихся трем классам деятельности учреждения. Получены оценки избыточности информации по монограммам, диграммам, триграммным условным энтропиям. Ниже исследуется эффективность предложенных повышению достоверности информации методов, которые учитывают вероятностные процессы, требуют использования средней

статистики ошибок и механизмов извлечения количественных характеристик статистической, естественной, структурно-технологической и лингвистической избыточности.

Таблица 1

Экспериментальные результаты оценки избыточности информации

Группа документов	Концепты	Природа информации	Размер в символ.	Условная энтропия (в бит)			Избыточность
				моно-граммы	ди-граммы	три-граммы	
I	1	текстовая	600	0,62	0,45	0,31	0,72
	2	алфавитно-цифровая	540	0,57	0,42	0,27	0,67
	3	цифровая	470	0,38	0,27	0,22	0,69
II	4	текстовая	712	0,71	0,46	0,32	0,38
	5	алфавитно-цифровая	615	0,61	0,33	0,21	0,49
	6	цифровая	621	0,63	0,46	0,33	0,75
III	7	текстовая	564	0,60	0,44	0,26	0,64
	8	алфавитно-цифровая	495	0,50	0,35	0,24	0,51
	9	цифровая	539	0,53	0,42	0,29	0,63

Метод повышения достоверности информации на основе кодового расстояния. Метод повышения достоверности информации основан на использовании концептуального принципа контроля заголовок протоколов ППП в телекоммуникационных сетях.

Особенностью метода является кодирование элемента концепта вводимого документа двоичными кодами, например Боуз – Чоудхури, Рид - Соломона, ASCII, применение которых требует определения кодового расстояния и нужного размера искусственной избыточности. Формируется проверочный шаблон в виде заголовка ППП. Производится контроль достоверности закодированной n - разрядным двоичным кодом информации на основе проверочного шаблона с размером N - максимальным числом двоичных символов.

Для обнаружения одно, двух, трёх и q - кратных ошибок вычисляется соответствующие кодовые расстояния, которые задаются в виде следующих результатов:

$$d = 1,\ N = 2^n;\ d = 2,\ N = 2^{n-1};$$

$$d = 3,\ N > 2^n/(1+n);$$

$$d = 2q + 1,\ N = 2^n/(1+n).$$

Вводимый шаблон принят как инструмент, предназначенный для обнаружения и исправления ошибок элементов вводимого документа. Представляет интерес методика определение общей метрики кодовых расстояний для шаблонного способа проверки достоверности информации.

Общая методика определения метрики кодовых расстояний. Для элемента концепта вводимого документа задаётся последовательность длиной N символов $x^N = (x_1, x_2, ..., x_N)$ и элементы набора проверочного концепта эталонного документа в виде последовательности $y^N = (y_1, y_2, ..., y_N)$.

Метрика расстояния для проверки достоверности информации задаётся в виде

$$d^N(x^N, y^N) = \frac{1}{N}\sum_{k=1}^{N} d(x_k, y_k).$$

Для разграничения достоверной и недостоверной части последовательности символов длиной N устанавливается некоторая пороговая величина d_p из множества значений D_1, $d_p \in D_1$.

Для обнаружения ошибок информации используется также M - подмножество элементов в виде набора - эталонного концепта документа, размещенного в контейнерах, как x^N и $\tilde{x}^N$. В соответствии с проверкой достоверности информации на основе набора – эталона фиксируется информационное поле пакета, правила контроля k^N, механизмы сравнения введенного документа f_N и эталонного ϕ_N.

Повышение достоверности информации на основе сравнения изображений вводимого и эталонного документов. Пусть f_N - контейнер изображения вводимого документа, который отображается последовательностью фрактальных элементов x^N и ϕ_N - контейнер - эталон, включающий последовательность модальных фракталов $\tilde{x}^N$.

Механизм обработки информации концепта с элементами m, правилами k^N запишется в виде:

$$f_N = x^N \times M \times k^N \to x^N;$$
$$x^N = f_N(\tilde{x}^N, m, k^N).$$

Область информации вводимого контейнера f_N ограничивается величиной D_1 в виде:

$$\sum_{\tilde{x}^N \in X^N}\sum_{k^N \in K^N}\sum_{m \in M}\frac{1}{|M|}p(\tilde{x}^N, k^N)d^N(\tilde{x}^N, f_N(\tilde{x}^N, m, k^N)) \le D_1,$$

где $p(\tilde{x}^N, k^N)$ - параметр, от значения которого зависит эффективность метода повышения достоверности информации.

Исследование проведено для равномерного и условного распределения

$Q^N(y^N/x^N)$. Нижняя граница оценки достоверности информации, оцениваемая по этой условной функции, запишется, как

$$\sum_{x^N \in \mathbf{X}^N} \sum_{y^N \in X^N} d^N(x^N, y^N) Q^N(y^N / x^N) p(x^N) \le D_2,$$

где D_2 - нижняя граница достоверной части информации, $D_2 \le D_1$.

Для достижения требуемого уровня достоверности информации в контейнере регулируется значение D_2.

Представляет интерес рассмотрение также случая, связанного с использованием усредненной вероятности ошибок информации в виде:

$$\sum_{m,k^N,\tilde{x}^N,y^N} d^N(\tilde{x}^N, y^N) Q^N(y^N / f_N(\tilde{x}^N, m, k^N)) p(\tilde{x}^N, k^N) \le D_2.$$

Оптимизация достоверности информации в контейнере изображений. Задача повышения достоверности информации связана с обеспечением робастности изображения вводимого документа, задаваемого качественной функцией f_N на основе механизма сравнения со изображением эталонного с качественной функцией $f*_N$. Условие для обеспечения робастности изображений документов задается, как

$$f_N(i,k,b) \approx f*_N(i+\sigma_n^2, k, b) \to \min,$$

где k,i,b - соответственно ключи, индекс контейнера изображения и специфических знаков вводимого документа;

σ_n^2 - погрешность оценки робастности изображений в контейнерах вводимого и эталонного документа.

Вводится оператор T, который модифицирует обработки контейнера изображения с учетом его функции плотности распределения вероятностей $w(\alpha)$- вводимого контейнера и $w^*(\beta)$ - эталонного контейнера.

Оператор T выбирается так, чтобы выполнялось следующее условие

$$T[(w(\beta/\alpha), R(S_0))] = T[(w(\alpha), R(S_w))] = T[(w^*(\beta), R^*(S_i))].$$

Оптимизация достоверности информации достигается благодаря механизму извлечения параметров, которые обусловливают использование различной природы избыточности информации $R(S_i)$ в структурах соответственно $R(S_0)$ - незаполненного контейнера, $R(S_w)$ - заполненного контейнера, $R^*(S_i)$ - модифицированного контейнера изображений.

Представляет интерес получение ответа на вопрос, что какое количество информации $I(\alpha,\beta)$. Ответ представляется механизмом

извлечения избыточности при использовании статистических, логических и семантических связей элементов и отношений концептов документа.

Количество информации связано с оценкой условной энтропии и функции распределения $w(\beta/\alpha)$, которые на плоскости задаются в виде:

$$I(\alpha,\beta) = R(S_0) \oplus R(S_w) \oplus R^*(S_w) w(\beta/\alpha) P_{\alpha\beta},$$

где $I(\alpha,\beta)$ - количество информации, учитывающее природу избыточности, используемой в контейнере изображения;

$\oplus$ - оператор суперпозиции.

$$P_{\alpha\beta} = \begin{cases} 0, & при \quad \alpha \neq \beta; \\ \dfrac{P}{B}, & при \quad \alpha \le \sigma_n^2 \le \beta; \\ 1, & при \quad \alpha = \beta, \end{cases}$$

$B = m^n$ - объем изменения информации, m - основание кода;

n - разрядность кода.

Многокритериальная оптимизация достоверности информации. Вероятность события, происходящих на сетевом уровне классической модели OSI обозначим в виде $P_{ПП}$ - правильный прием (ПП); P_H - необнаруженных ошибок (НО); P_o - обнаруженных ошибок (ОО) [8].

Вероятность обнаружения ошибок определяется, как

$$P_0 = 1 - (P_{ПП} + P_H),$$

где $P_{ПП} = (1 - p_c)^n$ - вероятность правильного приема элемента;

$p_c = \dfrac{P_{oik}}{n_c}$ – средняя вероятность ошибки элемента концепта с длиной n в предположении, что она равновероятная;

P_{oik} - вероятность ошибки одного элемента концепта документа.

Вероятность необнаруженной ошибки задаётся в виде условия

$$P_H < p_c n / 2^r; \; n_c p_c << 1,$$

где r - число разрядов двоичного кода элемента концепта документа.

Рассмотрим определение вероятностей правильного, искаженного приема и уничтожения ППП.

Вероятность правильного приема ППП $P_{ПП}$ на конечном узле маршрута передачи определяется, как

$$P_{ПП} = \prod_{i=1}^{N} P_{Пi},$$

где N - общее количество элементов в контейнере ППП; $P_{Пi}$ - вероятность

правильного приёма i - го элемента.

Алгоритм повышения достоверности информации на основе проверочного шаблона, связан со следующими ситуациями, когда [9]:

- ППП принимается правильно, но он зафиксирован на первом узле маршрута с вероятностью НО;

- ППП зафиксирован на втором узле, однако произошел ошибочный и правильный прием (ОПП).

Вероятность появления ошибок в контейнере передачи информации с N элементами и $N-1$ транзитными узлами маршрута запишется в виде:

$$P_{ИСК} = P_{НО} \cdot \sum_{k=1}^{N-1} P_{ПП}^{k-1} \cdot (P_{ПП} + P_{НО})^{N-k} .$$

Вероятность уничтожения ППП при N узлах запишется, как

$$P_{УС} = P_{ОО1} + \sum_{k=2}^{N} P_{ООk} \cdot \prod_{i=1}^{k-1} (P_{ППi} + P_{НОi}) .$$

Вероятности появления событий ПП, НО и ОО в каждом узле запишутся в виде:

$$P_{УС} = P_{ОО} \cdot \sum_{k=1}^{N} (P_{ПП} + P_{НО})^{k-1} .$$

Определено, что увеличение количества транзитных узлов в маршруте передачи ППП уменьшает вероятности его уничтожения. При этом вероятность передачи ППП с ошибочной информации увеличивается, однако ее значение будет на несколько порядков ниже, чем вероятности уничтожения и правильного приема ППП.

Ниже рассматривается метод, разработанный для повышения достоверности информации с применением механизмов кластеризации, которые имеют следующие особенности:

- каждый элемент концепта документа помещается в доступные классы и выбирается та конфигурация маршрута передачи ППП, в которой вероятность правильного приема максимальная;

- всякий раз, когда какой-то элемент концепта перемещается в новый класс, тогда не затрагиваются счетчики других классов;

- меняются только счетчики, которые используются при перемещении элемента концепта из класса c_i в класс c_k;

- последующие слова первоначально ищутся в классах слов, которые могут следовать за предыдущим и т.д.

Метод повышения достоверности информации с механизмом кластеризации. Особенность алгоритма метода заключается в следующем. Распознающее слово ищется в классе с наибольшим значением частоты их появления в предположение, что наиболее вероятное слово следует за проверяемым. Если оно там не находится, то поиск осуществляется в классе (классах) слов, которые могут стоять на первом месте в списке концептов.

Если оно и здесь не находится, тогда либо выдается отказ, либо поиск осуществляется по полному словарю словоформ. Разработана схема поиска слов с механизмом кластеризации [10].

Механизм обновления счетчиков реализуется по следующей схеме:

$$\forall w: N(c_j, w) = N(c_j, w) - N(w_i, w);$$
$$\forall w: N(c_k, w) = N(c_k, w) - N(w_i, w);$$
$$N(c_j) = N(c_j) - N(w_i);$$
$$N(c_k) = N(c_k) + N(w_i);$$
$$N(c_k, w) = \sum_{\forall i: w_i \in c_k} N(w_i, w),$$

где w_i - весовое значение элемента концепта w.

Реализация схемы с механизмом кластеризации также связана с формированием $(N_c - 1)$ частотного словаря словоформ, а все оставшиеся слова помещаются в N_c - м классе.

Для тестирования схемы сформированы классы, состоящих из 20, 50, 100, 200, 500 слов плюс к тому в каждый класс включаются следующие спецсимволы: $\langle b \rangle$ - начало предложения; $\langle e \rangle$ - конец предложения; $\langle n \rangle$ - число; $\langle u \rangle$ - неизвестное слово.

Размер словаря словоформ адаптируется в соответствие частотными словарями, состоящими из N_v слов, которые представляют массив информации с примерным объемом 64 Кб.

Установлено, что приблизительно 1000 наиболее часто используемых слов из словаря словоформ естественного языка задаётся распределением

$$\varsigma(s) = 1/n^s.$$

Когда экспонента $s > 1$, тогда использование приведенного распределения связано с трудностями вычислительного характера.

В связи с этим для отражения бесконечного множество слов $s > 1$ вводится следующая функция [11]:

$$\varsigma(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} < \infty.$$

Теперь распределение наиболее часто используемых слов по закону Парето запишется в виде

$$O(i) = \frac{m}{i^{\theta} H_{\theta}(V)},$$

где $1/i^{\theta}$ - частота появления элемента концепта в тестируемой последовательности; V – размер словаря словоформ; m – ранг к - грамм; $H_{\theta}(V)$ - гармоничное число θ-го порядка, значение которого колеблется между 1 и 2.

Эффективность механизма кластеризации, как правило, зависит от $(N-1)$ класса предшествующих транзитных слов, а в случае двухмерного распределения зависит лишь только от единственного класса. Вероятность принадлежности проверяемого слова к $(N-1)$ классу предшествующих слов оценивается по следующей логарифмической функцией

$$LL_{bi}(c) = \sum_{i=1}^{N_w} \log P(w_i \setminus C(w_{i-1})) = \sum_{j=1}^{N_c} \sum_{i=1}^{N_v} N(c_j, w_i) \cdot \log \frac{N(c_j, w_i)}{N(c_j)}.$$

Она упрощается в виде

$$LL_{bi}(c) = \sum_{j=1}^{N_c} \sum_{i=1}^{N_v} N(c_j, w_i) \cdot \log N(c_j, w_i) - \sum_{i=1}^{N_c} N(c_j) \cdot \log N(c_j).$$

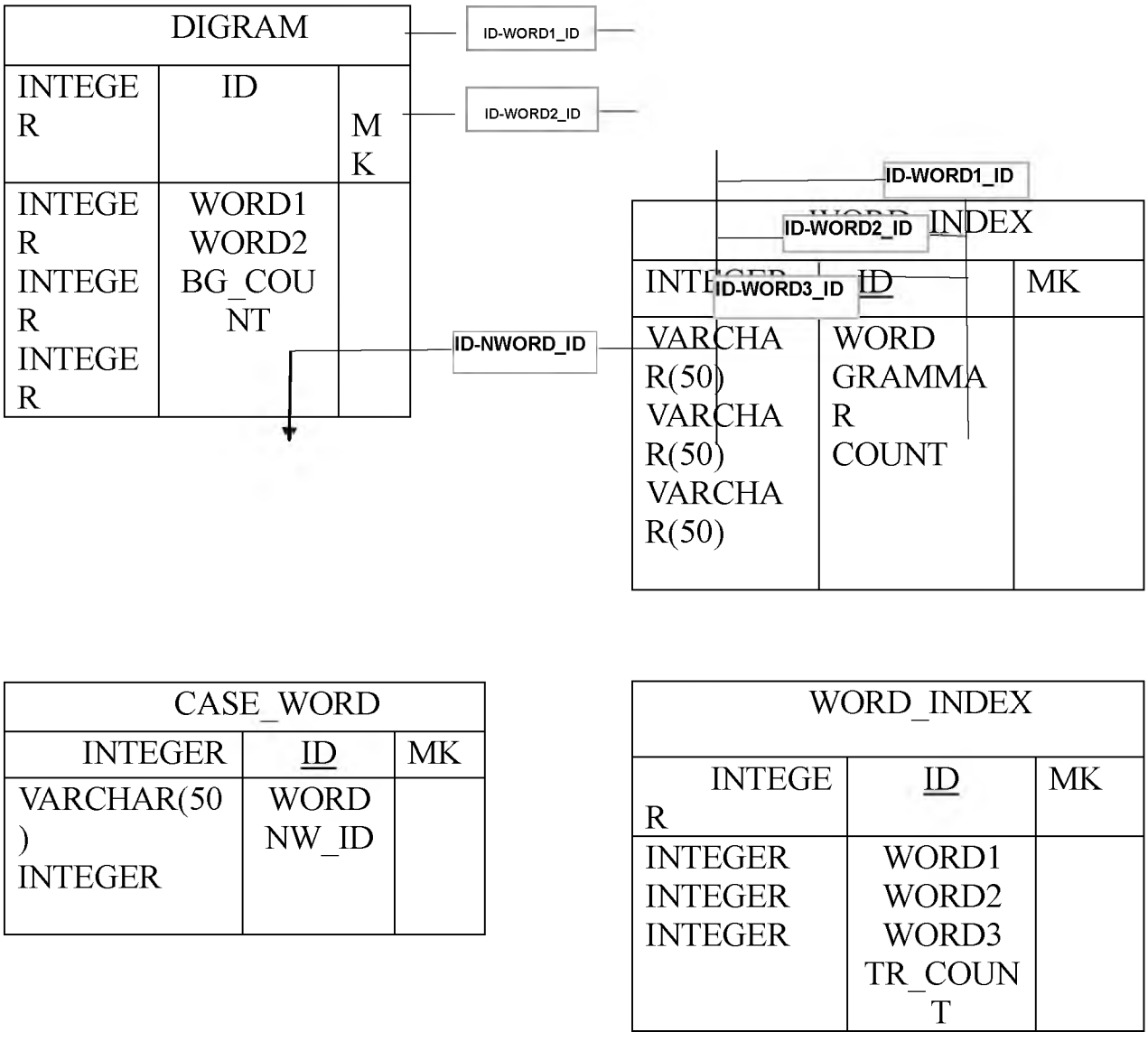


Рис.1. Схема повышения достоверности информации документа.

Следует отметить, что методы повышения достоверности информации документов связаны с многозначностью концептов, либо терминов и от вида инструментов распознавания, классификации, вычисления метрических расстояний и выявления набора «ссылок» из БД. В связи с этим разработана рациональная схема для реализации алгоритма повышения достоверности информации на основе n-грамм.

Реализация схемы повышения достоверности информации на основе n-грамм. Схема основана на реализации реляционной БД лингвистической информации, элементы которой представляются следующими таблицами: «DIGRAM» - диграммные исходы; «WORD_INDEX» - морфологическая информация; «CASE_WORD» - словарь словоформ; «TRIGRAM» - триграммные исходы.

На рис.1 проиллюстрирована схема повышения достоверности информации.

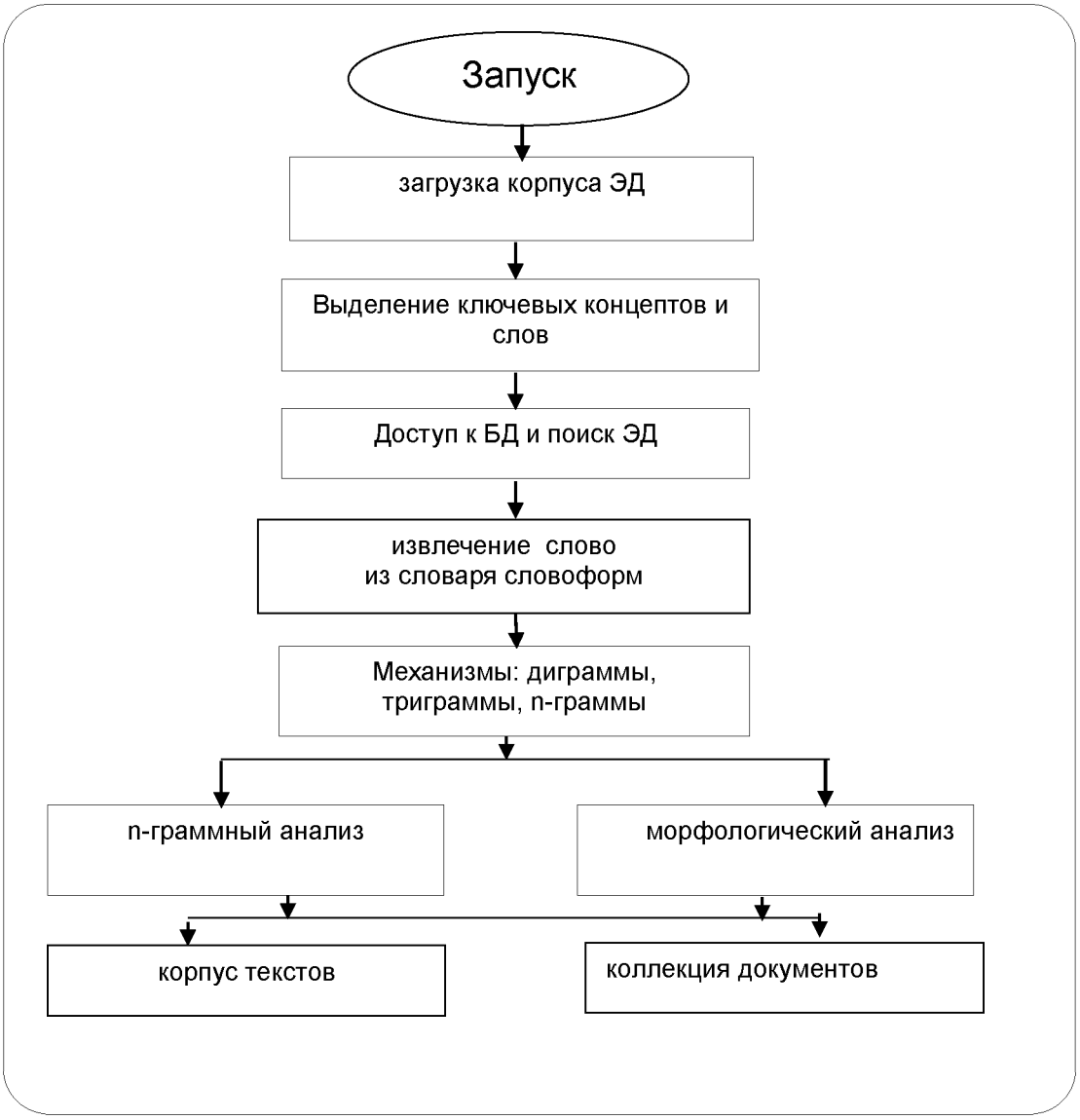
Для коррекции словарной базы словоформ выполняются следующие функции ввода и построения: начальных форм слов в таблицу «WORD_INDEX»; возможных словоформ в таблицу «CASE_WORD»; диграммной и триграммной частотных словарей; механизма контроля (МК) достоверности информации.

III. ЗАКЛЮЧЕНИЕ

Разработан программно-алгоритмический комплекс (ПАК) для повышения достоверности информации ЭД на языке программирования C#.NET, который включает разработанные алгоритмы повышения достоверности информации. Исследована коллекция документов делопроизводства, характерной деятельности ВУЗов. БД лингвистической информации ориентированы на реализации СУБД SQL Server 2008 по архитектурно трехступенчатому паттерну «клиент - сервер». Сформирован словарь словоформ из 670 тысячи слов; частотные словари из 75 тыс. диграмм и 46 тыс. триграмм.

Интерфейс ПАК имеет одно оконное приложение с шестью вкладками: «настройка»; «нормализация текста»; «снятие омонимии»; «словосочетания в тексте»; «анализ»; «ключевых слов».

На рис.2 проиллюстрирована схема взаимодействия программных модулей.



Разработан интерфейс ПАК, который имеет следующие преимущества:

- уменьшается ширина поисковой области, сокращается число нерелевантных документов;
- эффективность инструментов обработки информации проявляется наилучшим образом, трудоёмкость поиска снижается, а в замен огромного списка, представляются пользователю лишь только наборы тематических групп, которые позволяют игнорировать не интересующие документы.

ЛИТЕРАТУРА

[1] Белоногов Г.Г., Калинин Ю.П., Хорошилов А.А. Теория и практика построения систем автоматической обработки текстовой информации - М.: Русский мир, 2004. - 248 с.

[2] Архипова Н.И., Кульба В.В., Косяченко С.А., Чанхиева Ф.Ю. Исследование систем управления. - М.: ПРИОР, 2002. - 132 с.

[3] Хемминг Р. Теория кодирования и теория информации. – М.: Радио и связь, 1983. - 305 с.

[4] Блейхер Р. Теория и практика кодов, контролирующих ошибки. - М.: Мир, 1986. - 230 с.

[5] Белоногов Г.Г., Зеленков Ю.Г. Алгоритм автоматического обнаружения орфографических ошибок в текстах // М.: ВИНИТИ. – Москва, 1986. - 15 с.

[6] Ватолин Д.С. Алгоритмы сжатия изображений. – Москва: МГУ, 1999. - 130 с.

[7] Жуманов И.И. Разработка теории, исследование, практическое применение методов контроля и формирования информации со статистической избыточностью. – докт. дисс., Ташкент, УзНПО «Кибернетика». - 1983.

[8] Амирсаидов У.Б. Оценка достоверности и потери пакетов в сетях передачи данных \\ Вестник ТУИТ. - Ташкент, 2009. - №2. - с 73-77.

[9] Жуманов И.И., Ахатов А.Р. Оценки достоверности передачи изображений элементов текста в телекоммуникационных сетях// Журнал «Химическая технология. Контроль и управление» - Ташкент, 2010 - № 2. - с. 23-30.

[10] 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.

[11] Jumanov I.I., Akhatov A.R. Fuzzy Semantic Hypernet for Information Authenticity Controlling in Electronic Document Circulation Systems // 4-th International Conference on Application of Information and Communication Technologies, 12-14 october 2010, Section 2, IEEE. - Tashkent, 2010. - p.21-25.