Chemical Technology, Control and Management

Volume 2019 | Issue 3 Article 6

6-29-2019

An improvement of information reliability for electronic documents based on knowledge bases with implication matrix adjustment

I.I. Jumanov
Samarkand State University, Address: University blv. 15, 140104,Samarkand city, Samarkand region, Uzbekistan, Phone: (+99866)2293558, olimjondi@mail.ru

Kh.B. Karshiev
Samarkand State University, Address: University blv. 15, 140104,Samarkand city, Samarkand region, Uzbekistan, Phone: +998936805400(м.)., xusan2005@mail.ru

An improvement of information reliability for electronic documents based on knowledge bases with implication matrix adjustment

Cover Page Footnote

Tashkent State Technical University, SSC «UZSTROYMATERIALY», SSC «UZKIMYOSANOAT», JV «SOVPLASTITAL», Agency on Intellectual Property of the Republic of Uzbekistan

ISSN 1815-4840
CHEMICAL TECHNOLOGY. CONTROL AND MANAGEMENT
2019, №3 (87) pp.49-57
International scientific and technical journal
journal homepage: https://uzjournals.edu.uz/ijctcm/
Since 2005

УДК 519.681.5

AN IMPROVEMENT OF INFORMATION RELIABILITY FOR ELECTRONIC DOCUMENTS BASED ON KNOWLEDGE BASES WITH IMPLICATION MATRIX ADJUSTMENT

I.I.Jumanov[1], Kh.B.Karshiev[2]

[1,2]*Samarkand State University*
Address: University blv. 15, 140104,Samarkand city, Samarkand region, Uzbekistan
E-mail: [1]olimjondi@mail.ru, Phone: (+99866)2293558; [2]xusan2005@mail.ru, Phone: +998936805400(м.)

Abstract: *The new methods, algorithms and software tools have been developed to increase the reliability of information based on the extraction and use of quantitative, qualitative, specific characteristics, useful properties (interrelations between elements of document concept relationships). New approaches and principles to improve the traditional techniques and developed new methods of processing documents using a database (DB) and setup knowledge base (KB) to optimize the accuracy of the information have been proposed. The algorithms of increase of reliability of elements of the concept of the input document based on the proximity measures and comparison with the elements of the test set – model, and mechanisms for setting BZ binary matrix implications. Tools of structural model application, extraction of interelement connections, compression of description space, use of statistical, dynamic, specific characteristics and features of the document for detection and correction of the distorted information taking into account parameters of influence of external environment are designed. The effectiveness of the algorithms was investigated by the criteria of reliability, complexity and cost of information processing.*

Keywords: *information accuracy, segmentation, interconnectedness of elements, relation of concepts, a set of standards, implication matrix, knowledge base, tuning, generalized algorithm.*

Аннотация: *Ҳужжатларнинг сонли, сифат, махсус таснифлари, фойдали хоссаси ва элементларини ўзаро боғланиши, концептлар муносабатини ажратиш ва фойдаланишга асосланган ахборот ишончлилигини ошириш муаммоси талқин этилган ҳамда уларнинг усуллари, алгоритм ва дастурий воситалари яратилган. Анъанавий технологияни такомиллаштиришга ёндашув ва принциплар таклиф этилган ва ахборот ишончлилигини оширишни оптималлаштириш учун маълумотлар базаси, билимлар базасини мослаштириш механизмларидан фойдаланувчи ҳужжатларга ишлов беришнинг янги усуллари ишлаб чиқилган. Кирувчи ҳужжат концепти элементлари ишончлилигини яқинлик ўлчови ва назорат қилувчи эталон – жамлама элементлари билан солиштиришга асосланган алгоритмлар ҳамда иккилик импликацияли матрицага эга ББ мослаштириш механизми ишлаб чиқилган. Ташқи муҳит таъсири параметрини инобатга олиб, ахборот хатолигини аниқлаш ва таҳрирлаш учун ҳужжатнинг таркибий моделини қўллаш, элементлар боғланишини ажратиш, тасниф фазосини қисиш, статистик, динамик, хусусий таснифи ва хоссаларидан фойдаланувчи воситалар лойиҳалаштирилган. Алгоритмлар самарадорлиги ахборотга ишлов беришнинг ишончлилиги, мураккаблиги ва харажатлар қиймати мезонлари бўйича тадқиқ қилинган.*

Таянч сўзлар: *ахборотларнинг ишончлилиги, сегментация, элементларнинг ўзаро боғланиши, концептлар муносабати, жамлама – эталон, импликациялаш матрицаси, билимлар базаси, созлаш, умумий алгоритм.*

Аннотация: *Сформулирована проблема, а также разработаны методы, алгоритмы, программные средства повышения достоверности информации, основанные на извлечении и использование количественных, качественных специфических характеристик, полезных свойств и взаимосвязей между элементами, отношений концептов документов. Предложены подходы и принципы совершенствования традиционной технологии и разработаны новые методы обработки документов с использованием базы данных (БД) и механизмов настройки базы знаний (БЗ) для оптимизации достоверности информации. Разработаны алгоритмы повышения достоверности элементов концепта вводимого документа на основе меры близости и сравнения с элементами проверочного набора – эталона, а также механизмов настройки БЗ с двоичной матрицей импликации. Спроектированы инструменты применения структурной модели, извлечениее межэлементных связей, сжатие пространства описания, использование статистических, динамических, специфических характеристик и*

особенностей документа для обнаружения и коррекции искаженной информации с учетом параметров влияния внешней среды. Эффективность алгоритмов исследована по критериям достоверности, трудоёмкости и стоимости обработки информации.

Ключевые слова: *достоверность информации, сегментация, взаимосвязанность элементов, отношение концептов, набор –эталон, матрица импликации, база знаний, настройка, обобщенный алгоритм*

Введение

Существующие технологии электронного документооборота производственно - технологических комплексов управления основаны на использовании расширенной априорной информации, сведений о свойствах, особенностях и характеристиках данных. Кроме того, методы повышения достоверности информации становятся пригодными только лишь в условиях, когда задается параметрическое описание изображения документа и отсутствует неопределенность [1-3].

Однако документы в системах характеризуются структурной сложностью, разнотипностью, многосвязанностью концептов (элементов, признаков, атрибутов), значительным их числом, вариациями динамики, наличием различного рода информационной избыточности и неопределенности в параметрах, ограниченностью априорных сведений, а также неточностью обработки информации.

Настоящая работа посвящена разработке методов, алгоритмов, программных инструментариев для повышения достоверности информации, основанных на использовании механизмов извлечения количественных, качественных специфических характеристик, полезных свойств, взаимосвязей между элементами и отношений концептов документов. Исследование предполагает совершенствование традиционных и разработку новых методов обработки документов на основе использования базы данных (БД) и базы знаний (БЗ) с механизмами настройки переменных.

Подходы и принципы оптимизации достоверности информации

Предложен принцип использования ключевых концептов документа с существующими элементами и сведениями, отражаемых в виде слов, типовых текстов, массивов, наборов переменных, которые предоставляют возможности для повышения достоверности информации на основе ведения проверочных специализированных карт, набор–эталона. Проверочный набор-эталон, который содержит формализованные структурные части документа, обладает высокой чувствительностью к изменениям значений элементов, учитывает межэлементные связи, силу отношений концептов при проверке информации документов [4-7].

Реализация этого принципа должна быть направлена на уменьшение размерности поискового пространства документа, которое обусловлено трудоемким процессом обработки информации с целью устранения проблематичности поиска комбинаторного типа, происходящей в традиционной технологии на основе просмотра n элементов при $\frac{n!}{2n}$ вариантах [6].

В традиционных технологиях извлечениа знаний, полезных сведений, необходимых характеристик из документов выполняются вручную, либо полуавтоматическими методами и такое обстоятельство при большом числе документов приводит к значительному росту числа обращений к БД и БЗ, что отрицательно влияет на экономическую эффективность методов обработки информации [7].

В этой связи предложен подход, требующий формирование и применение БД и БЗ с механизмами контроля достоверности элементов и регулирование переменных документов, в результате которых предоставляются возможности для получения малоитеративных алгоритмов обработки данных. Особенностью задач оптимизации достоверности информации на основе настройки БЗ является реализация матрицы импликации, правил поиска, инструментов использования связей между элементами и отношений концептов, проверочного

набор–эталона, включающего в структуре специфические, статистические, динамические характеристики, текстурные особенности и свойства документов [7,8].

Для оптимизации достоверности информации представляют большую значимость результаты исследования и разработки следующих механизмов:

- извлечения и использования межэлементных связей и их структурной модели, сжатие размера пространства описания документов и устранение неинформативных элементов;
- использования статистических параметров, динамических, специфических характеристик документов, скрытых и полезных свойств, закономерностей данных и элементов документа;
- регулирования параметров, формирования оптимального проверочного набор - эталона, адекватного описания документа;
- обнаружения, корректировки искаженных элементов документа в условиях влияния внешней среды;
- совмещения их возможностей и построения обобщенного алгоритма обработки информации.

Для оптимизации обработки документов применяются механизмы сегментации, кластеризации, выделения фрагментов, фрактальных характеристик, текстурных особенностей изображений образа, ключевых концептов из документов, которые одновременно используются при повышении достоверности информации.

Рассмотрим алгоритмы повышения достоверности информации, которые основываются на механизмах формирования БЗ с матрицами двоичных импликаций, использования инструментов поиска, связанности элементов и отношений концептов и проверочного набор - эталона.

Алгоритмы повышения достоверности информации на основе проверочного набор–эталона

Концепт с элементами однородного документа, как правило, имеет мало различий от элементов такого же концепта, включенного в проверочный набор – эталон, что способствует обнаружению и коррекции искаженной информации. В результате первой итерации алгоритма формируется набор элементов, относящихся к универсуму значений элементов только лишь проверочного концепта, которые считаются достоверными. Не относящиеся элементы к универсуму значения элементов проверочного концепта набора – эталона считаются недостоверными. Эффективность повышения достоверности информации оценивается по пороговому значению в пределах границ функции расстояния, которое разделяет последовательность элементов концепта на два подмножества: достоверной и недостоверной информации с целью их различия друг от друга. Изложим правила контроля достоверности элементов ключевого концепта документа на основе набор–эталона [9, 10].

Пусть проверочный набор–эталон задан из элементов ключевого концепта документа, которые сегментируются путем образования групп, принадлежащих подмножеству $x_{p1}, x_{p2},...., x_{pm_p} \in X_p$, $p = \overline{1.r}$, где x_{pi} - концепт документа, задаваемый N - мерным вектором элементов $x_{pi} = x_{pi}^1, x_{pi}^2,....., x_{pi}^N, i = \overline{1, m_p}$.

Функция расстояния $\theta_i(x_{pj}, x_{pk})$, $k = \overline{1, m_p}$, $j = \overline{1, m_p}$, $j \neq k$ между элементами вводимого документа и проверочными элементами набор – эталона задается в виде

$$\theta_i(x_{pj}, x_{pk}) = \begin{cases} 1 & если\ (x_{pj}, x_{pk}) = 0, i = \overline{1.N}; \\ 0 & если\ (x_{pj}, x_{pk}) \neq 0, i = \overline{1.N}, \end{cases}$$

где x_{pj}, x_{pk} - соответственно проверяемый x_{pj} и проверочный x_{pk} элементы, задаваемые в пространстве булевых символов в сегменте X_p.

Первое условие «1» выражает наличие сходства двух объектов т.е наличие «достоверной информации», второе же «0» означает наличие между элементами несходства и «недостоверной информации».

Величина расхождения отражает различие произвольного l-го концепта документа от проверочного концепта в пространстве булевых символов; она задаётся в виде [11]

$$\Gamma_l(x_{pj}, x_{pk}) = \sum_{k=1}^{m_p}\sum_{j=1}^{m_p}\theta_1(x_{pj}, x_{pk}).$$

Оценка достоверности информации элемента l-го концепта на основе набор–эталона задаётся в виде

$$\Gamma_l(x_{pj}, x_{pk}) = \sum_{k=1}^{m_p}\sum_{j=1}^{m_p}\rho_l(x_{pj}, x_{pk}),\ j = \overline{1.m_p}; k = \overline{1.m_p};\ j \neq k.$$

где $\rho_l(x_{pj}, x_{pk})$ - мера близости элементов концепта вводимого документа и элементов проверочного концепта, включенных в структуру набор – эталона.

Мера близости элементов концепта вводимого документа с учетом весов связей ($w^1, w^2, ..., w^N$) на элементы проверочного концепта, включенных в структуру набор – эталона определяется в виде [12]

$$\Gamma_w(w, x_{pk}) = \sum_{k=1}^{m_p}\sum_{j=1}^{m_p}\rho_l(w, x_{pk}),$$

Проверяется неравенство,

$$\Gamma_w(w, x_{pk}) > \Gamma_w(w, x_{pj}),$$

и если оно выполняется то i- й элемент проверяемого концепта считается «достоверной информацией».

Оценка достоверности информации на основе проверочного набор–эталона

В качестве критерия оценки достоверности информации ЭД выступает вероятность ошибок элемента в структуре концепта документа; последняя определяется, исхода из следующего закона распределения [13]:

$$P_i = c_n^i p^i (1-p)^{n-i,},$$

где p - вероятность появления поэлементной ошибки; $(1-p)^{n-i}$ - вероятность отсутствия ошибок в структуре концепта; n - длина последовательности закодированных элементов концепта; $i-$ кратность ошибок (одно, двух, трёх, n- кратных).

Для оценки достоверности элементов концепта с двухкратной ошибкой учитываются следующие условия, когда: ошибка произойдет в двух элементах концепта; сумма вероятностей двух ошибочных элементов останется равной сумме вероятностей двух верных элементов; знак суммы ошибочных элементов не изменится. Пример: когда элементы концепта документа кодируются десятичным кодом с длиной *n=12*, применяется следующая оценка достоверности информации

$$P_н = c_{12}^2 p^2 (1-p)^{10}.$$

Предложенные выражения вероятностей использованы при исследовании эффективности алгоритмов повышения достоверности информации на основе проверочного набора–эталона, основанных на использовании взаимосвязанности элементов и отношений концептов документа.

Представляет также интерес оценка в общем виде следующей вероятности ошибки p_c одной единицы информации (бит, элемент текста) в рассматриваемом концепте длиной n в предположении, что информация является независимой и равновероятной

$$p_c = \frac{P_{oik}}{n},$$

где P_{oik} - вероятность искажения элементов концепта документа.

Если элементы концепта являются верными, то такая вероятность задаётся, как

$$P_в = (1 - p_c)^n .$$

Вероятность не обнаружения ошибочного элемента концепта задаётся, как

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

где r - число элементов в концепте документа.

Вероятность обнаружения ошибочного элемента концепта задаётся, как

$$P_0 = 1 - (P_в + P_н) .$$

Алгоритм решения задачи состоит из следующих шагов:

Шаг 1. Объекты набор - эталона вносятся в БД, которая формируется в разрезе всех $\rho = \overline{1,r}$ элементов X_p класса сегментов.

Шаг 2. Вычисление функции расстояния в пространстве булевых символов. Определение различия элементов проверочного концепта набор-эталона от элементов концепта вводимого документа.

Шаг 3. Вычисление и формирование функции расстояния в пространстве булевых символов, используемых при оценке достоверных элементов концепта документа.

Шаг 4. Оценка величины, отражающая различие произвольного *j*-го концепта документа от концепта набор – эталона.

Шаг 5. Оценка вклада механизма проверке достоверности *j*-го концепта в коллекции документов.

Шаг 6. Оценка результата контроля достоверности вводимого документа на основе решающего правило набора - эталона.

Алгоритм протестирован с механизмами настройке БЗ с переменными P_i, $i = 1,..,8$ и X_j, $j = 1,..,26$. Получено:

- наборы, проверяемых элементов концепта вводимого документа

$$\|T_i\| = \begin{pmatrix} 11110001100001111100111110 \\ ... \\ 00100100011111101011111100 \end{pmatrix};$$

-набор – эталона $T_{эт} = (11110101100001100110111110)$;

- результаты проверки достоверности элементов концепта

$$T_1 : 1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,0\,0\,0\,1\,1\,1\,1\,1\,0$$

$$\ldots . \qquad\qquad \ldots .$$

$$T_8 : 0\,0\,1\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,1\,0\,1\,0\,0\,0\,1\,1\,1\,1\,0\,1$$

Результат запуска алгоритма для повышения достоверности информации на основе набор–эталона задаётся числом голосов, равным $\Gamma_w(w, x_{pk}) = d$ - расстоянию по Хеммингу, значение которого отражается в столбцах и строках матрицы.

Алгоритмы повышения достоверности информации на основе БЗ с матрицами импликаций

Повышение эффективности механизма определения различия между элементом концепта вводимого документа и элементом концепта проверочного набора–эталона выполняется путем ведения матрицы импликаций U, в каждой строке которой фиксируются выявленные признаковые различия. Матрица импликаций U является целочисленной, элементы столбцов которой сопоставляется со элементами столбцов некоторой проверочной матрицы Q, а элементы строки сопоставляются с всевозможными парами переменных v, l, $v \in \{1,2,...,\sigma(Q^a)\}$, $l \in \{1,2,...,\sigma(Q^b)\}$, обозначенных через a и b.

Переменные $\sigma(Q^a)$ и $\sigma(Q^b)$ определяются количеством элементов в строке подматрицы Q^a либо подматрицы Q^b в проверочной матрице Q.

Различие $u_{i,j}$ j-го элемента, $j \in \{1,2,...,m\}$ в строке подматриц Q^a и Q^b задается в виде $u_{i,j} = \left| q^a_{v,j} - q^b_{l,j} \right|$, где $q^a_{v,j}$ - элемент v строки матрицы импликации и j- й элемент проверочной подматрицы Q^a; $q^b_{l,j}$ - элемент l строки матрицы импликации и j-й элемент проверочной подматрицы Q^b.

Значимость различия $u_{i,j}$ j-го элемента задается следующим правилом

$$u_{lj} = \begin{cases} 1, & \text{если } q^a_{i,j}; \\ 0, & \text{если } q^b_{l,j}, \end{cases}$$

где «1» и «0» – обозначения величины значимости различия $u_{i,j}$.

Величина различия концепта вводимого документа и проверочных подматриц Q^a и Q^b элементов a, b $(a \neq b)$ вычисляется в виде

$$\delta^j_{a,b} = \left| q_{a,j} - q_{b,j} \right|,$$

где $\delta^j_{a,b}$ - различие элементов a и b; q_{aj} - значимость различия для элемента a; $q_{b,j}$ - значимость различия для элемента b.

Матрицы импликации иллюстрируют:

- представления знаний, поскольку одной строкой проверочной матрицы Q задаются значения элементов в интервальной форме;
- представления различия любых пар из заданного множества;
- представления исходной матрицы T с двоичными элементами, столбцы которой сопоставляются со столбцом проверочной матрицы Q.

Методика проектирования механизма настройки БЗ с матрицей импликации

Механизм настройки БЗ с матрицей импликации задаётся кортежами

$$M = F < D, S, M_0 >,$$

где M - модель, представляющая некоторую функцию, возвращающая БЗ; D - набор данных; S - правила настройки переменных; M_0 - начальная (пустая) структура БЗ, не содержащей никаких правил.

Модель формирования набора данных D представляется кортежами

$$D = F < A, T, W >,$$

где $A = \{a_1,...,a_i,...,a_m\}$ - набор независимых элементов a_i; $T = \{a_{m+1},...,a_k,...,a_t\}$ - множество целевых концептов; $W = \{w_1,...,w_r,...,w_z\}$ - множество векторов данных; $w_r = \{v_{1,r},...,v_{i,r},...,v_{m,r},v_{m+1,r},...,v_{k,r},...,v_{m+t,r}\}$ - набор элементов $v_{i,r} \in D(a_i)$ со значениями a_i, соответствующих концепту документа.

Модель возвращения БЗ задается кортежами

$$M = F < R, K >,$$

где $R = (r_1,...,r_i,...,r_n)$ – множество правил извлечения свойств данных; $K = (k_1,...,k_p,...,k_g)$ – количество векторов; $k_p = \left| (w_r : w_r \in W,\ v_{r,t} = v_{r,t}) \right|$ - элементы вектора целевого концепта a_t, имеющие значение $v_{r,p}$ в виде $v_{r,1} \cup ... \cup v_{r,p} \cup ... \cup v_{r,z} = D(r)$.

БЗ после запуска принимает одно из следующих состояний: M_0 – начальное состояние; M_1 – состояние после 1-го изменения; M – состояние возвращенной модели.

Модель с настройкой представляется набором пар переменных в виде

$$S = ((s_1, p_1),...,(s_y, p_y),...,(s_z, p_z))$$

где s_y, p_y – соответствующая пара для y-ой переменной модели.

Алгоритм повышения достоверности информации выполняется на основе набор-эталона, имитирующего последовательно выполняющихся операций в виде

$$F = \{O_1,...,O_p,...,O_q\},$$

где O_p – операция, изменяющая БЗ из одного состояния в другое.

Операции настройки БЗ задаются данными D, настройкой S и моделью M_i. Результат настройки БЗ в виде модели M_{i+1}, приводящая изменению его состояния задается, как

$$M_{i+1} = O < D,S,M_i >.$$

В целом настройки БЗ представляются комплексом операции в виде

$$F = \{O_1(D,S,M_0),...,O_p(D,S,M_p),...,O_q(D,S,M_q)\}.$$

Кроме того, механизм настройки БЗ включает условные операторы, циклы, в том числе циклы по векторам и по данным.

Условный оператор получения модели M_{i+1} БЗ представляется в виде

$$M_{i+1} = D < D,S,M_i >= \{ d(D,S,M_i),O_t(D,S,M_i),O_f(D,S,M_i)\},$$

где $d(D,S,M_i)$ – условная функция, возвращающая сигналов «достоверная» или «ошибочная» информация на основании модели M_i; $O_t(D,S,M_i)$ – операция, выполняющая возвращение функции $d(D,S,M_i)$ при сигнале «достоверная информация»; $O_f(D,S,M_i)$ – операция, выполняющая возвращение функции $d(D,S,M_i)$ при сигнале «недостоверная информация».

Цикл работы механизма по модели M_{i+1} представляется в виде

$$M_{i+1} = C < D,S,M_i >= \{ d(D,S,M_i),O_c(D,S,M_i)\},$$

где $O_c(D,S,M_i)$ – операция, выполняющаяся по функции $d(D,S,M_i)$ возвращение сигнала «достоверная информация».

Упорядоченная последовательность функциональных блоков механизма настройки БЗ для модели M_q представляется в виде

$$M_q = F < M_0,S,D >= (b_0(D,S,M_0),...,b_k(D,S,M_k),...,b_q(D,S,M_{q-1})).$$

Совмещение инструментов обработки информации

Построен обобщенный алгоритм в структуру которого включены следующие функциональные модули: формирования двоичной матрицы импликации U; нахождения соответствующих характеристических элементов, концептов документа; формирования строк матрицы Q, упорядочивание строк R' матриц Q, U' и U; вычисления весовых коэффициентов элементов; формирования двоичной матрицы U'' путем замены значений всех элементов, отличных от «0» на значение «1»; формирования покрытий столбцов из матрицы U'' и проведения тестирования.

Разработан комплекс программ на базе обобщенного алгоритма с использованием средств Builder C++. Архитектура комплекса открытая и представляет иерархическую структуру соединения в системе.

Главный модуль – резидентный, имеет встроенную систему команд, выполняет функции ядра. Все остальные модули являются динамически подключаемыми и подразделяются на модули системных данных и пользовательского интерфейса. Первый модуль предназначен для работы с БД и БЗ. Входными данными являются структура БЗ, объекты базы знаний, название элементов и номера концептов. Второй модуль предназначен для выбора элементов покрытия столбцов из матрицы U''. Третий модуль формирует номера целочисленных характеристических элементов, номера классификационных концептов, структуру БЗ, номера тестирующих набора – эталона документов.

В результате запуска комплекса формируется выходные результаты, которые представляют вещественный вектор с весовыми значениями элементов концепта, а также двоичная матрица U'', являющаяся исходной.

Исследование проведено по функции контроля исполнения организационно – распорядительных документов (ОРД) в деятельности Самаркандского государственного университета.

Анализ результатов исследования

Для тестирования программного комплекса на основе обобщенного алгоритма выделены 6 ключевых концептов из 40 однородных документов, каждый концепт включает 352 элементов, которые разбиты на 5 сегментов. При этом размер матрицы двоичной импликации Q равен 352x40, а размер матрицы R равен 352x1. Количество характеристических элементов концепта документа в тестирующем наборе - эталона документа равно 165. При этом размер матрицы Q равен 381x165.

Исследования проведены с целью обнаружения и коррекции ошибок различных типов. Установлено закономерность частотных искажений (монограмм, диграмм, триграмм и *n*-грамм) и определено, что наиболее вероятными являются однократные, двухкратные и трехкратные ошибки.

В первом исследовании размер матрицы U' равен 381x2 (762), а во втором – размер матрицы U'' равен также 381x2 (762). Число тестирующего набора – эталона по каждому концептов документа равно 6.

В первом исследовании построена матрица U'' размером 27352x165, а также выявлены следующие факты: 155 элементов являются ошибочными; 134 элементов статистически и логически связаны с 139 элементами; 30 признаков статистически связаны с 28 элементами; 23 концептов логически связаны с 31 концептами документа. Исправления обнаруженных ошибок за счет правил алгоритмов, использующих статистических связей позволили сократить пространства с ожидаемыми искажениями информации до 161 элементов концепта на 94 %. А алгоритмы, использующих логических связей между концептами позволяет исправлять ошибки в элементах концепта документа до 98 %.

Во втором исследовании построена матрица U'' размером 26149x165 и выявлены следующие факты: 155 элементов является ошибочными; 30 элементов статистически связаны с 28 элементами; 23 элементов логически связаны с 31 элементами концепта документа. Определено, что показатель вероятности необнаруженных ошибок традиционной технологии находится в пределах $10^{-4} - 10^{-5}$, а показатель временных затрат на повышение достоверности информации 100 документов с общим объемом информации 10^8 элементов (десят. знак) за один цикл работы равен 2,4 часа.

Установлено, что программный комплекс позволяет сокращать значения трудоёмкости обработки информации с 4,5 часов до 1,1 часа. Коэффициент выигрыша в достоверности

информации увеличивается в 3-5 раз, что выше показателя достоверности традиционной технологии, основанной на использовании корректирующих кодов.

Механизмы использования логико-семантических свойств элементов концептов и структурно - технологической избыточности данных документа повышают достоверность информации до трех порядков. Обобщенный алгоритм способствует повышению коэффициента обнаружения ошибочных элементов в документе с 88% до 98,7 %, при этом временные и стоимостные затраты на обработку документов сокращаются в 7-8 раза.

References:

1. Kogalovskij M.R. Perspektivnye tekhnologii informacionnyh sistem. M.: DMK Press: Kompaniya AjTi, 2003. –288c.
2. Cisco Visual Networking Index: Forecast and Methodology, 2011 - ¬2016. May 30, 2012, sajt - URL:
3. Horoshko M.B. Modifikaciya algoritma bulevogo poiska / M. B. Horoshko // Izvestiya vysshih uchebnyh zavedenij. Severo-Kavkazskij region. Seriya: Tekhnicheskie nauki. - 2011 № 3 - S. 14-18.
4. Gavrilova T. A., Horoshevskij V. F. Bazy znanij intellektual'nyh sistem. SPb: Piter, 2000. –384 s.
5. Ermakov A.E. Izvlechenie znanij iz teksta i ih obrabotka: sostoyanie i perspektivy // Informacionnye tekhnologii, 2009, № 7, –s. 50-55.
6. Hubaev G.N. Algoritm sravneniya slozhnyh sistem po kriteriyu funkcional'noj polnoty // Materialy konferencii «Ekonomiko- organizacionnye problemy analiza, proektirovaniya i primeneniya informacionnyh sistem»/RGEA.- Rostov n/D, 2007.
7. Jumanov I.I., Akhatov A.R. The control of information transfer reliability in intellectual control systems on the basis of statistical redundancy //Sixth World Conference on Intellectual Systems for Industrial Automation, Uzbekistan, TSTU. - Tashkent, 2010. – p. 70-75.
8. ZHumanov I.I., Ahatov A.R. Intellektual'nyj kontrol' dostovernosti tekstovoj informacii na osnove ucheta svojstv i raspredelenij dannyh // «Himicheskaya tekhnologiya. Kontrol' i upravlenie» - TGTU, - Tashkent, 2011. - № 1 (37), - s.41-47.
9. Voroncov K.V. Obzor sovremennyh metodov po probleme kachestva obucheniya algoritmov // Tavricheskij vestnik informatiki i matematiki, 2004. №1. -S. 5-25.
10. Zagorujko N.G., Kutnenko O.A., Zyryanov A.O. i dr. Obuchenie raspoznavaniyu obrazov bez pereobucheniya // Mashinnoe obuchenie i analiz dannyh. 2014, t. 1, №7.-S.891-901.
11. Kamilov M.M., Nishanov A.H., Beglerbekov R.ZH. Primenenie reshayushchego pravila dlya vybora informativnyh naborov priznakov // Himicheskaya tekhnologiya. Kontrol' i upravlenie. - Tashkent, 2017, №3. - 82-85.
12. Fazylov SH.H.. Nishanov A.H., Mamatov N.S. Metody i algoritmy vybora informativnyh priznakov na osnove evristicheskih kriteriev informativnosti, Tashkent: «Fan va texnologiya». 2017 g. -132s.
13. Jumanov I.I., Ahatov A.R., Tishlikov S.A. Adaptivnye algoritmy optimizacii processov obnaruzheniya i ispravleniya oshibok v sisteme obrabotki tekstov // «Vestnik TUIT», Tashkentskij universitet informacionnyh tekhnologij. - Tashkent, 2011. - №3/2011. - s. 31-37.