ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
INFORMATICS AND INFORMATION TECHNOLOGIES

УДК 519.681.5

ПОВЫШЕНИЕ ДОСТОВЕРНОСТИ ИНФОРМАЦИИ СИСТЕМ МОНИТОРИНГА ОРГАНИЗАЦИОННО – РАСПОРЯДИТЕЛЬНОЙ ДЕЯТЕЛЬНОСТИ

Жуманов И.И., Каршиев Х.Б.

Разработаны методы повышения достоверности информации с механизмами определения рационального размера наборов-эталона, интервала принадлежности достоверных элементов, редукции избыточных элементов, семантического поиска документа, извлечения статистических, специфических характеристик информации.

Предложены правила регулирования количества сегментов, левой и правой границ сегментов, размера набор–эталона элементов документа для минимизации общей вероятности необнаруженных ошибок.

Исследована эффективность алгоритма повышения достоверности информации на основе определения оптимальных границ разделения достоверной и недостоверной части информации по функциям условного распределения.

Разработан и реализован программный комплекс повышения достоверности информации с модулями предварительной обработки документов с выполнением поиска, распознавания, кластеризации и сегментации, настройки границ сегментов элементов документа, проверки достоверности информации по эквивалентности элементов и взаимной эквивалентности сегментов и наборов.

Ключевые слова: электронный документ, набор – эталон, достоверность информации, информационная избыточность, взаимосвязанность элементов, отношение концептов, кластеризация, программный комплекс

Эталон–жамламанинг қулай ўлчамини, ишончга эга элементлар тегишлилик интервалини, ортиқча бўлган элементларни редукцияловчи, ҳужжатни семантик равишда қидирувчи, маълумотлардан статистик, хусусий таснифларини ажратувчи механизмларга эга бўлган ахборот ишончлигини ошириш усуллари ишлаб чиқилган. Хатоларни аниқламаслик умумий эҳтимолини минималлаштир учун ҳужжат элементи сегментлари сонини, сегмент чап ва ўнг чегараларини, эталон – жамлама қулай ўлчамини мувофиқлаштирувчи қоидалар таклиф этилган. Шартли тақсимот функцияси бўйича маълумотларнинг ишончли ва ишончсиз бўлган қисмларини

бўлакловчи мақбул чегараларни аниқлаш асосида ахборот ишончлилигини ошириш алгоритми самарадорлиги тадқиқ қилинган. Ҳужжатларга дастлабки ишлов беришда қидириш, таниш, кластерлаш ва сегментлаш амалларини бажарувчи, ҳужжат элементи сегментлари чегараларини мувофиқлаштирувчи, элементлар эквивалентлиги ҳамда сегмент ва жамламаларнинг ўзаро эквивалентлигини текшириш модулларига эга бўлган ахборот ишончлигини оширувчи дастурий мажмуа ишлаб чиқилган ва жорийлаштирилган.

Таянч сўзлар: электрон ҳужжат, информация ишончлилиги, эталон – жамлама, информация ортиқчалиги, элементлар боғликлиги, концептлар муносабати, кластерлаш, дастурий мажмуа

In this article have been developed methods for increasing the information reliability with mechanisms for determining the rational size of a set of standard, the interval of belonging of reliable elements, reduction of redundant elements, semantic search for a document, extraction of statistical, specific characteristics of information. The rules for regulating the number of segments, the left and right borders of the segments, the size of the set of the standard elements of the document to minimize the overall probability of undetected errors are proposed. The effectiveness of the algorithm for increasing the reliability of information have been investigated on the basis of determining the optimal boundaries for separating reliable and unreliable parts of information by conditional distribution functions. Developed and implemented a software package to increase the information reliability with the modules of document preprocessing with the search, recognition, clustering and segmentation, setting the boundaries of segments of document elements, checking the accuracy of information on the equivalence of elements and mutual equivalence of segments and sets.

Keywords: electronic document, information accuracy, set of standards, information redundancy, interconnectedness of elements, relation of concepts, clustering, software package

I. ВВЕДЕНИЕ

Актуальность темы. В автоматизированных системах управления производственно–технологическими комплексами ключевой задачей исследования считается разработка и применение моделей и методов, способствующих достижению высокой достоверности информации, оптимизации обработки документов производства при ограниченных априорных сведениях, параметрической неопределенности и низкой точности обработки данных [1,2].

Проблематичность применения методов для повышения достоверности информации в системах электронного документооборота (СЭД)

обуславловлена трудоемкими вычислениями, высоко итеративными алгоритмами и затратами, связанными с обнаружением и исправлением ошибок в документах[3].

Большую теоретическую и практическую значимость имеют методы, основанные на обобщении свойств и использовании особенностей типичных инструментов поиска, распознавания, формирования базы данных (БД) и базы знаний (БЗ) при существенно меньших временных затратах на переработки документов. При этом, реализация механизмов извлечения и использования статистических параметров, специфических, динамических характеристик, скрытых закономерностей, полезных знаний в документах существенно повышает возможности рассматриваемых методов [4].

Настоящая работа посвящена разработке методов и алгоритмов повышения достоверности информации на основе определения рационального размера набора–эталона; механизмов настройки переменных, в частности интервала значений элементов, концепта, параметров документа; механизма редукции избыточных элементов, признаков, параметров в структуре сети семантического поиска; механизма извлечения статистических, специфических характеристик, динамических свойств информации, особенностей элементов документа и формирования БД и БЗ с настройкой переменных.

II. ОСНОВНАЯ ЧАСТЬ

Построение алгоритмов повышения достоверности информации. Пусть заданы наборы, которые выделены из символьного пространства элементов по каждому концепту, параметру документов, определены правила отбора информативных элементов, а также условия для повышения достоверности информации в режиме реального времени.

Элемент концепта документа обозначен через x, а проверочный набор - эталон через S. Набор включает x_j^s элементов, j - порядковый номер, который он принадлежит $j = 1,2,...,N$; s - порядковый номер элемента концепта набора, где $s = 1,2,...,S$.

Каждая последовательность элементов концепта x_j^s отражается элементами y_j^s набора - эталона.

Общая последовательность возможных элементов документа разбивается на последовательность участков (сегментов).

Для запуска алгоритма определяется число разделяющих плоскостей перпендикулярно оси каждого элемента, количество сегментов в общей последовательности, порядковые номера сегментов, левая и правая границы сегментов элементов концепта, рациональный размер проверочного набора.

Алгоритм повышения достоверности информации на основе набора - эталона. Алгоритм представляется следующими шагами [5].

Шаг 1. Инициализация параметров набора.

Вектор входного элемента x задается в виде матрицы, в строках которой выполняется линеаризация, а столбцы формируют массив $y^* = \{y^s\}$ с бинарными элементами.

Шаг 1.1. Формирование массива $\{D_j\}$ с размером, равным количеству элементов N. Для каждого концепта и документа определяется число сегментов в общей последовательности элементов.

Шаг 1.2. Установить: $D_j = 0$, $j = 1,2,...,N$, где j - порядковый номер текущего элемента. Фиксируется набор в виде переменной S.

Шаг 1.3. Установить порядковый номер текущего элемента начиная с $i = 1$.

Шаг 2. Если $i \le N$, то перейти на шаг 3. Иначе на шаг 11.

Шаг 3. В буфер массива x концепта по i-му документу занести набор $x(j) = x_i^s$.

Шаг 3.1. В буфер массива y занести копию массива $y^* : y(s) = y^s$.

Шаг 4. Провести сортировку элементов в массивах x и y.

Шаг 4.1. Формировать массивы элементов x и y в порядке возрастания их номеров.

Шаг 4.2. Фиксируется порядковый номер набора, $s = 1$.

Шаг 4.3. Если $s \le S$, то перейти на шаг 4.4, иначе на шаг 5.

Шаг 4.4. Устанавливается порядковый номер набора $k = s + 1$.

Шаг 4.5. Если $k \le S$, то перейти на шаг 4.6, иначе на шаг 4.8.

Шаг 4.6. Если $x(s) > x(k)$, то установить следующие переменные:

$$z = x(s);\ x(s) = x(k);\ x(k) = z;\ z = y(s);\ y(s) = y(k);\ y(k) = z.$$

Шаг 4.7. Установить: $k = k + 1$ и перейти на шаг 4.5.

Шаг 4.8. Установить: $s = s + 1$ и перейти на шаг 4.3.

Шаг 5. Установить: $s = 1$, $k = 1$.

Шаг 6. Если $s \le S$, то для хранения в буфере массивов k-го сегмента из общей последовательности элементов устанавливается $a^t = x(s)$ и перейти на шаг 7. Иначе перейти на шаг 11.

Шаг 7. Если $s < S$, то устанавливается $y(s) = y(s+1)$.

Шаг 8. Если $s = S$, то устанавливается $y(s) = y(s-1)$.

Шаг 8.1. Фиксируется сегмент $K(i,k)$ из общей последовательности элементов, k - порядковый номер сегмента; i-ый элемент внутри k-го сегмента.

Шаг 8.2. Устанавливаются $A(i,k)$ и $B(i,k)$, соответственно левая и

правая границы сегмента. Перейти на шаг 10.

Шаг 9. Если $s < S$ и $y(s) \neq y(s+1)$, то устанавливаются следующие:

$$K(i,k) = y(s);\ A(i,k) = a^t;\ B(i,k) = x(s);\ k = k+1;\ s = s+1.$$

Перейти на шаг 6.

Шаг 10. Установить: $i = i+1$, перейти на шаг 2.

Шаг 11. Останов.

Разработан обобщенный алгоритм повышения достоверности информации, в котором синтезируется процедуры выделения сегментов, определения границ сегментов в общей последовательности элементов для каждого концепта, производится отбор информативных элементов, формируется рациональный проверочный набор с механизмом адаптации параметров документа.

Повышение достоверности информации на основе проверки эквивалентности элементов и сегментов в последовательности. Исследование направленно на разработку механизма сокращения количества сегментов в общей последовательности элементов документа. Чем меньше количество сегментов, тем рациональнее проверочный набор.

Идентифицируется следующие переменные:

$\{x, y\}$ – наборы пар, которые представляют массивы информации;

$\{D_j\}$ – количество сегментов;

$\{A(i,k)\}$, $\{B(i,k)\}$ – нижняя и верхняя границы сегментов $\{K(i,k)\}$;

$\{K(q)\}$ – порядковые номера каждого сегмента.

По результатам обработки информации проверяются, во - первых, эквивалентность сегмента элементов концепта вводимого документа $[A(i,k); B(i,k)]$ сегменту элементов набор - эталона $[A(j,q); B(j,q)]$. Во - вторых, проверяется эквивалентность элемента x_k вводимого документа и элемента в набор - эталоне x_q.

Коэффициент эквивалентности k-го сегмента i-го концепта, q-му сегменту j-го концепта g-го набора элементов документа определяется исходя из следующих условий:

$$n(x_i^s, x_j^g, k, q) = \begin{cases} 0, \text{ если } K(i,k) \neq K(j,q); \\ 0, \text{ если } B(i,k) < x_i^s \text{ или } x_i^s < A(i,k); \\ 0, \text{ если } B(j,q) < x_j^g \text{ или } x_j^g < A(j,q); \\ 1, \text{ если } K(i,k) = K(j,q), A(i,k) \le x_i^s \le B(i,k), A(j,q) \le x_j^g \le B(j,q), \end{cases}$$

где $s = 1,2,...,S$; $g = 1,2,...,S$; $i = 1,2,...,N$; $j = 1,2,...,N$; $k = 1,2,...,k_i$; $q = 1,2,...,k_q$.

Коэффициент эквивалентности элемента концепта вводимого документа и элементу концепта набор – эталона определяется, как

$$N(i,k,j,q)=\sum_{s=1}^{S}\sum_{\substack{g=1,\\ g\neq s}}^{S} n(x_i^s,x_j^g,k,q). \qquad (1)$$

Взаимная эквивалентность набор – эталонов задается в виде:

$$E_{i,k,j,q}=\min\left\{\frac{N(i,k,j,q)}{N_{i,k}},\frac{N(i,k,j,q)}{N_{j,q}}\right\}=\frac{N(i,k,j,q)}{\min\{N_{i,k},N_{j,q}\}}.$$

Взаимная эквивалентность i-го элемента, k-го сегмента вводимого концепта j-ым элементом, q-го сегмента в набор – эталоне определяется, как

$$E_{i,j}=\frac{\sum_{k=1}^{k_i}\sum_{q=1}^{k_j}E_{i,k,j,q}}{\max\{k_i,k_j\}}. \qquad (2)$$

Определение рационального числа сегментов. Алгоритм включает следующие шаги.

Шаг 1. Инициализация. Формирование набора пар $x=\{x_i\}$ и $y=\{y_s\}$, $i=1,2,...,s$, $s=1,2,...,S$.

Шаг 2. Вычислить характеристики наборов.

Шаг 2.1. Найти параметры: $A(i,k)$, $B(i,k)$, $K(i,k)$, $N_{i,k}$, k_i.

Шаг 2.2. Определить коэффициенты: $N(i,k,j,q)$, $E_{i,k,j,q}$, $E_{i,j}$.

Шаг 3. Установить: $i=N$.

Шаг 4. Если $i>1$, то выполнять шаги 4.1 и 4.2.

Шаг 4.1. Когда $\forall j$, $j\neq i$, $j=1,2,...,(i-1)$ и $E_{i,j}=1$, тогда удалить x_i и установить $N=N-1$.

Шаг 4.2. Установить: $i=i+1$. Перейти на шаг 4.

Шаг 5. Установить: $i=N$.

Шаг 6. Если $i\geq 1$, то выполнять шаги 6.1 и 6.2.

Шаг 6.1. Установить: $k=k_i$.

Шаг 6.2. Если $k\geq 1$, то выполнять шаги 6.2.1 - 6.2.3, иначе перейти на шаг 7.

Шаг 6.2.1. Рассчитать:

$$c=\sum_{j=1}^{N-1}\sum_{q=1}^{k_j}E_{i,k,j,q},E_{i,k,j,q}=1.$$

Шаг 6.2.2. Если $c\geq 1$, то удалить k-й сегмент i-го элемента в общей

последовательности элементов документа.

Шаг 6.2.3. Установить: $k_i = k_i - 1$. Перейти на шаг 6.2.

Шаг 7. Останов.

Оптимизация достоверности информации на основе настройки границ сегментов. Для настройки устанавливаются нижняя и верхняя границы сегмента в виде:

$$\text{если } A(i,k) \le \mu_{i,k}(x_i) \le B(i,k), \text{ то } y_i^s = K(i,k),$$

где y_i^s - порядковый номер сегмента в s-м набор – эталоне по элементу i; $\mu_{i,k}(x_i)$ – степень принадлежности i-го элемента в установленные границы.

Задаются функции плотности распределения (ФПР) вероятностей элементов, определяются границы, общий интервал для последовательности элементов концепта документа, порог разделения последовательности элементов на достоверных и недостоверных. Повышение достоверности информации обеспечивается правилами проверки по каждому набор - эталону соответствия (принадлежности) элементов вводимого концепта в пределы разрешенных границ. Так, например, соответствие элемента вводимого концепта $\mu^0(x^s)$ и элемента концепта $\mu^1(x^s)$ набор – эталона требует выполнения следующего условия:

$$\mu^0(x^s) = \max \mu_{i,k}(x_i); \quad \mu^1(x^s) = \max \mu_{i,k}(x_i).$$

Следовательно, достоверность элементов вводимого документа по набор - эталону проверяется по коэффициенту эквивалентности, как

$$K(i,k) = \begin{cases} 1, & \text{если} \quad \mu^1(x^s) > \mu^0(x^s); \\ g(x_{i,k}), & \text{если} \quad A(i,k) \le \mu_{i,k}(x_i) \le B(i,k); \\ 0, & \text{если} \quad \mu^1(x^s) \le \mu^0(x^s). \end{cases} \qquad (3)$$

Случай «1» означает максимальное соответствие элемента вводимого документа элементу набор – эталону с коэффициентом $K(i,k) = 1$; Случай «0» означает полное несоответствие элементов с коэффициентом $K(i,k) = 0$;

Случай «$g(x_{i,k})$» означает соответствие элементов с коэффициентом, вычисляемым по формуле (1). Аналогично осуществляется процедура повышения достоверности информации по коэффициенту взаимной эквивалентности сегментов, концептов, наборов и даже однородных документов.

Алгоритм повышения достоверности информации по условным функциям принадлежности. Пусть элементы концепта вводимого документа представляются множеством $\{\alpha_i^1\}$ с функцией распределения вероятностей (функцией принадлежности) $W(\alpha_i)$ и $W(\beta_j)$ с функцией распределения вероятностей элементов набор – эталона [6,7].

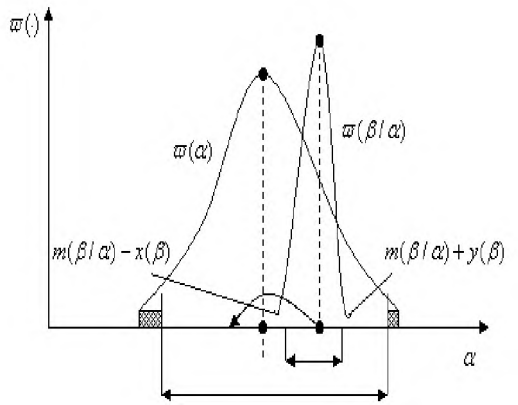

Рис. 1.Контроль информации.

Фиксируется $\{\alpha_p^{1,k}\}$ – разрешенное подмножество, которое определяется нижней - $x(\alpha)$ и верхней - $y(\alpha)$ границами, а также $\{\alpha_з^{1,k}\}$ – запрещенное подмножество, которое отражает вероятность не принадлежности элементов в эти границы. Если элемент α принадлежит подмножеству $\{\alpha_p^{1,k}\}$, то информация считается достоверной. Если элемент α принадлежит подмножеству запрещенных значений $\{\alpha_з^{1,k}\}$, то информация считается ошибочной. Алгоритм повышения достоверности информации оставляет необнаруженными ошибки двух родов. Ошибки первого рода – «пропуск ошибок» возникает, когда элемент искажен, попадает в пределы разрешенного подмножество и алгоритмом считается достоверным. Ошибки второго рода – «ложная тревога», возникают, когда элемент правильный, находятся в запрещенном подмножестве, а алгоритмом считается недостоверным.

На рис.1, проиллюстрированы кривые функции распределения вероятностей $\varpi(\alpha)$ вводимого элемента α и элемента β проверочного набора $\varpi(\beta/\alpha)$ относительно элемента α [6].

Статистические связи между элементами α и β можно рассматривать как двумерную регрессионную зависимость, что позволяет наглядно проиллюстрировать эквивалентность элементов вводимого концепта и проверочного набора.

Общее решение задачи. Решение задачи заключается в нахождении оптимальных границ разделения достоверной и недостоверной информации по функции условного распределения с помощью механизма регулирования, который способствует достижению минимизации общей вероятности необнаруженных ошибок. В соответствии с гипотезой о равномерности распределения вероятностей ошибок P запишем

$$P = P_1 + P_2 = \frac{P}{B}\int_{x(\alpha)}^{y(\alpha)} \varpi(\beta/\alpha)d\alpha + (1-P)(1-\int_{x(\alpha)}^{y(\alpha)} \varpi(\beta/\alpha))d\alpha) \quad . \qquad (4)$$

Вероятность необнаруженных ошибок алгоритма зависит от условий передачи информации P, границ контроля достоверности информации, видов функций распределения вероятностей элементов, B - общего диапазона возможной принадлежности элементов.

Частное решение задачи. Требуется определение границ функций распределения вероятностей.

Рассматривается табличная функция условного распределения вероятностей элемента z с точностью до членов порядка $N=5$, которая аппроксимируется по выражению [7]:

$$F(z)=F(z)-\frac{1}{\sqrt{m}}\left(\frac{a_3}{3!}\right)F^{(3)}(z)+\frac{1}{m}\left[\frac{1}{4!}э_4F^{(4)}(z)+\frac{10}{6!}a_3^2F^{(6)}(z)\right], \quad (5)$$

где $F(z)$ - функция нормального распределения вероятностей с нулевым средним μ_1 и единичной дисперсией μ_2;

$F^{(\cdot)}(z)$ - ее производные;

a_3 и $э_4$ - асимметрия и эксцесса задаваемые, как $a_3=\frac{\mu_1}{\sqrt{\mu_2^3}}$;

$э_4=\frac{\mu_1}{\mu_2^2}-3$.

Нижняя и верхняя границы (пороги) z_1 и z_2 находятся численным решением нелинейных уравнений:

$$z=F(z_1)=0{,}5q;$$
$$F(z_2)=1-0{,}5q,$$

где q-уровень значимости $0{,}01\div0{,}1$;

$F(x)$- условная функция плотности распределения вероятностей $W(\beta/\alpha)$;

$z=\frac{\sum_{i=1}^{N}\alpha_i}{\sqrt{N_{\mu_2}}}$ - удовлетворяет одному из следующих условий $z>z_2$; $z<z_1$.

Движением по оси z сначала от z_1, а потом z_2 находятся их значения, которые дают более точные решения рассматриваемого уравнения.

По первому условию считается, что элемент соответствует пределам границ условной функции распределения и достоверным, а по второму условию считается недостоверным.

III. ЗАКЛЮЧЕНИЕ

По резудьтатам исследований разработан программно – алгоритмический комплекс (ПАК) повышения достоверности информации. ПАК включает следующие функциональные модули [8]:

– предварительная обработка информации с выполнением механизмов поиска, распознавания, кластеризации и сегментации;

– определение и настройка границ сегментов элемента концепта либо документа в информативном интервале редукция избыточных сегментов;

– определение и настройка рационального размера проверочного набора, коэффициентов эквивалентности, взаимной эквивалентности элементов сегмента, концепта, документа и проверочных наборов;

– оптимизация параметров ФПР элементов документа.

Разработанный обобщенный алгоритм повышения достоверности информации протестирован в среде пакета MATLAB.

Исследована эффективность механизмов сегментирования элементов концепта, адаптации параметров. Полученные результаты сопоставлены с эффективностью традиционной технологии контроля достоверности документов в системах управления.

Определено, что результаты позволяют спроектировать высокоточные инструменты анализа, обработки информации, адаптивного контроля достоверности информации документов в режиме реального времени.

Алгоритмы имеют прозрачные структуры, устойчивы к ошибкам и устраняют недостатки, характерные для высоко итеративных алгоритмов. Эффективно эмулируются свойства документов, используются механизмы формирования БД и БЗ, которые способствуют повышению достоверности информации до требуемого уровня и системы становятся менее затратными.

ЛИТЕРАТУРА

[1] Абдуллаев Д.А., Амирсаидов У.Б. Комплексная модель физического и канального уровней сети передачи данных \\ Вестник ТУИТ. - Ташкент, 2007. - №4. - с. 19-43.

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

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

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

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

[6] Жуманов И.И., Ахатов А.Р. Анализ качества передачи информации// «Вопросы кибернетики»: Сб. научн. тр. - Ташкент: ИК АН РУз, 2002. - №163. - с. 61-66.

[7] 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.

[8] Jumanov I.I., Akhatov A.R. Estimation of reliability for the system of mistakes dynamiic control at transfer and processing of the text information// Abstracts of Plenary and Invited Lectures of International School and Conference on Foliations, Dynamical Systems, Singularity Theory and Perverse Sheaves, 6-21 October 2009, SamSU, Uzbekistan. – Samarkand, 2010. – 80-85 p.p.