УДК 519.681.5

ПОВЫШЕНИЕ КАЧЕСТВА ФУНКЦИОНИРОВАНИЯ СИСТЕМ ЭЛЕКТРОННОГО ДОКУМЕНТООБОРОТА НА ОСНОВЕ МЕХАНИЗМОВ ИСПОЛЬЗОВАНИЯ СЕМАНТИЧЕСКОЙ ИЗБЫТОЧНОСТИ

Жуманов И.И.[1], *Каршиев Х.Б.*[1]
[1] Самаркандский государственный университет имени Шарофа Рашидова, Самарканд, Узбекистан
xusan2005@mail.ru

Аннотация. *Разработаны методические основы решения задач оптимизации поиска, хранения, обработки информации по критериям достоверности, трудоемкости и стоимости. Предложены оценки времени и стоимости ввода, передачи, хранения, обработки, обмена документами, обнаружения и исправления ошибок информации на основе механизмов использования семантической избыточности реализованных и лексографического синтеза структуры документа. Исследована эффективность механизмов использования статистических, логических, семантических и структурно – технологических связей элементов документов. Разработана и реализована вычислительная схема решения оптимизационной задачи на основе применения адаптивных методов стохастического случайного поиска, моделирования усеченной цепью Маркова и динамического программирования. Реализован программный комплекс повышения достоверности информации на основе использования адаптивного случайного поиска, сегментации и лексикологического синтеза структуры.*

Ключевые слова: *эффективность, система электронного документооборота, трудоемкость и стоимость обработки информации, достоверность информации, оптимизация, стратегия оптимизации, стохастическая модель.*

I. ВВЕДЕНИЕ

В системах электронного документооборота предприятий, учреждений и директивных органов, вводятся, передаются, хранятся, организационно – распорядительные документы (ОРД), в составе которых происходят искажения информации [1,2].

Для обеспечения целостности, сохранности, релевантности документов и достоверности информации решаются задачи оптимизации обработки электронных документов (ЭД) в соответствии с критериями минимизации объема документов в базах данных (БД), времени и стоимости получения достоверной ЭД из других информационных систем (ИС) [3-6].

Обеспечение качества обработки ОРД зависят от эффективности решения задач повышение достоверности информации на следующих этапах ИС [4-6]:

- ввода человек - оператором, техническими средствами сканирования и распознавания;
- передачи по: локальной, корпоративной, глобальной сети;
- хранения оперативной, среднесрочной, долгосрочной информации, формирования БД, базы знаний (БЗ), а также документопотоков из других источников;
- сервисного обслуживания, представления услуг пользователю с результатами.

Настоящая работа посвящена исследованию методических основ обеспечения качества функционирования СЭД на основе разработки и применения методов, моделей и алгоритмов повышения достоверности информации, основанных на использование избыточностей различной природы, в частности, статистической, логической, семантической и структурно – технологической.

II. ОСНОВНАЯ ЧАСТЬ

2.1. Основные подходы и принципы повышения качества функционирования СЭД.

В исследование, в качестве критериев эффективности функционирования систем приняты временный, стоимостный и вероятностный показатель – достоверная обработка информации ЭД, значения которых оценивается, начиная с ввода до представления выходного документа пользователю с раскладкой долевого вклада результативных переменных на каждом этапе функционирования СЭД [7].

Решение задачи оптимизация достоверности обработки информации документов начнем

с задачи размещения потоков ЭД в БД, минимизации времени и стоимости их представления пользователю из БД и других ИС, а также минимизации общего объема памяти, занимаемого документами системы.

Предложены соблюдения и применения следующих принципов при оптимизации [8]:

- минимизации времени и стоимости получения требуемого документа пользователем, которое складывается из времени передачи запроса, поиска документа, обработки информации, а также времени представления ЭД пользователю. Причем, целевая функция оптимизации зависит от способов организации и размещения ЭД в БД и в других ИС;

- минимизации времени и стоимости обмена документами пользователем и БД системы, а также между другими ИС;

- минимизации памяти для хранения ЭД в БД системы с целью снижения непериодических затрат на хранение, обновление, внесение изменений в документах во всех хранилищах ИС;

- минимизации стоимости получения нужного документа пользователем, передачи информации, повышения достоверности информации и коррекции содержимого документа.

Конструктивный подход к оптимизации размещения ЭД. Решения задач размещения ЭД в БД, занимаемых множеством элементами, атрибутами, концептами, характеристиками и другими данными остро востребованы. Они используются также при анализе и оценке качества функционирования СЭД.

Предложен новый подход, направленный на оптимизации размещения ЭД большого количества. Исследованы разнообразность, применяемых форматов ЭД - кадров их изображения, типичные инструменты поиска, распознавания, классификации [9].

Обоснованы эффективность предложенных механизмов формирования и использования шаблонов, эталонов, фреймов, вероятностных, логических, семантических свойств и характеристик информации ЭД [10].

Исследованы методы решения задачи нахождения локальных и глобального экстремумов целевой функции [11].

Рассмотрим подход к оптимизации размещения документов в БД на основе модели оптимального «раскроя», которая представляет NP-трудную задачу.

Предложены принципы модификации технологии ее решения, особенностями которых являются следующие.

Прямоугольный раскрой. Пусть задано поле - бесконечная полоса с шириной W, на котором размещаются m прямоугольных предметов (l_i, w_i), $i = 1,2,...,m$ - длина и ширина формата изображения документа, задаваемые при следующих условиях:

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

- при размещении на полосе, никакие два предмета не пересекаются друг с другом т.е

$$((x_i \geq x_j + l_j) \vee (x_j \geq x_i + l_i) \vee$$
$$(y_i \geq y_j + w_j) \vee (y_j \geq y_i + w_i))$$

для $i, j = 1,2,...,m, i \neq j$;

- никакой предмет не пересекает границ полосы т.е $(x_i \geq 0) \wedge (y_i \geq 0) \wedge (y_i + w_i \leq W)$ для $i = 1,2,...,m$. Геометрический смысл переменных W изложен в [2].

Требуется разместить на бесконечной полосе набор прямоугольных предметов так, чтобы занятая ими часть полосы была минимальной по длине. Следовательно, находится такой набор (x_i, y_i), чтобы

$$L = \max(x_i + l_i) \to \min,$$

где $L(l_i, w_i), (x_i, y_i)$, $i = 1,2,...,m$.

Раскрой круглых предметов. В данном варианте решения задачи размещения ЭД в БД, требуемая память для которых представляется на бесконечной полосе с шириной W для размещения, m круглых предметов, радиусы которых известны как r_i. Задача оптимизации решается при следующих условиях:

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

- при размещении на полосе никакие два предмета не пересекаются друг с другом т.е.

$$(x_i - x_j)^2 + (y_i - y_j)^2 \geq (r_i + r_j)^2$$
для $i, j = 1,2,...,m, i \neq j$;

- никакой предмет не пересекает границ полосы т.е.

$$(x_i - r_j \geq 0) \wedge (y_i - r_i \geq 0) \wedge (y_i + r_i \leq W).$$

Решениями задачи являются размещение на бесконечной полосе, набор круглых предметов минимальной длины. Размещение по возможности должно быт плотное. Оптимизирована длина полосы, занятая размещенными объектами.

Оптимизация времени и стоимости представления ЭД пользователю. Целевую функцию F1 определим в виде

$$F_1 = t' + t'' \to \min, \qquad (1)$$

где t' - среднее время получения пользователем ЭД из БД;

t'' - среднее время получения пользователем ЭД из других ИС.

Решения задачи оптимизации проведены в зависимости от применяемых механизмов размещения данных в БД, поиска, а также повышения достоверности информации [12].

Показатели трудоемкости (время) и стоимости обработки информации ЭД в системе оптимизируется по следующему функционалу

$$F_2 = \sum_{i=1}^{N}\sum_{j=1}^{n_i} W_{ij}^{'} + \sum_{i=1}^{M}\sum_{j=1}^{m_i} W_{ij}^{''} \to \min, \qquad (2)$$

где n_i - количество ЭД, хранимое в БД H_i, m_i - количество ЭД, хранимое в других ИС I_i, N - количество БД системы, M - количество других ИС, $W_{ij}^{'}$ - время либо стоимость представления услуг пользователю по документу d_j из БД H_i, $W_{ij}^{''}$ - стоимость либо время представления услуг по документу d_j из других ИС I_i. ЭД представляется в виде вложенного файла, набора файлов произвольного размера.

2.2. Механизмы оптимизации поиска объектов.

Для решения задачи оптимизации поиска нужного документа или информационной части документа предложен подход, направленный применению метода маршрутизации с требованием к нахождению оптимального решения задачи большой размерности.

Для оптимизации маршрута поисковых объектов использован принцип формирования специальной матрицы. В матрице поисковые объекты размещаются по строке, а по столбцу отображается последовательность правил, извлекаемых в маршруте [11].

В позиции (x,i) матрицы стоит 1 в том случае, когда объект x занимает i-е место в маршруте. В случае n поисковых объектов, сталкиваемся с $\frac{n!}{2n}$ различными правилами, извлекаемых в маршруте поиска и требуется найти среди них наиболее кратчайший.

Традиционный подход, направленный на оптимизации идентификации маршрута поисковых объектов модифицирован на основе подхода, связанного с применением нейронной сети (НС), в частности, сети Хопфильда. При этом, составляется матрица нейронов размера $n \times n$, которые взаимодействуют по строкам и столбцам.

В реализованном механизме оптимизации каждый нейрон обозначается двумя индексами x и i. Причем x отражает объект, а индекс i позицию в маршруте поиска, т.е. z_{xi} – это выход нейрона, в котором объект x размещен на i-й позиции маршрута.

НС Хопфильда оценивается вычислительной сложностью - n^4, n – размерность задачи и много итеративного подбора переменных, которые выполняются при подборе подходящей функции активации нейронов и архитектуры сети.

Для упрощения вычислений, связанных с поиском объекта предложено применение рекуррентной НС, которая позволяет снизить сложность маршрутизации с $O(n^4)$ до $O(n^2)$.

В работе предложен подход, направленный созданию механизма поиска на основе сочетания графевой модели. Такая совмещенная модель поиска объекта состоит из ребер, соединяющих ближайшие вершины.

Задача заключается в выборе новых вершин с оценкой вероятностей, зависящих от расстояний удаления от текущей. Чем оно имеет большее значение, тем меньше будет значение этой вероятности. Алгоритм модифицированного механизма поиска имеет следующие особенности:

- агент, запущенный для прохождения пути по маршруту поиска, сохраняет предыдущий маршрут поиска объекта;
- если агент уже был в этой вершины, то он само уничтожается;
- количество запросов соответствует выполнению следующего условия

$$H \cdot N \geq |G|,$$

где H - размер ЭД (количество элементов, признаков, концептов и др.), N - размер концепта документа (количество элементов), G - мощность БД (количество ЭД).

Минимизация непериодических временных и стоимостных затрат. Минимизация непериодических временных и стоимостных за-

трат на хранение, обновление, внесения изменений во все хранилище информации задается следующей целевой функцией F_3

$$F_3 = \sum_{i=1}^{N}\sum_{j=1}^{n_i} v_{ij} \to \min, \qquad (3)$$

где v_{ij} - объем информации ЭД d_j в памяти БД H_i.

Интегральный стоимостный критерий S_{int} задается в виде

$$S_{\text{int}} = (S_{xr} + S_{pxr} + S_{p\inf}) \to \min, \qquad (4)$$

где S_{xr} - стоимость хранения ЭД в БД системы, S_{pxr} - стоимость получения ЭД пользователем из других ИС, $S_{p\inf}$ - стоимость получения ЭД пользователем из БД системы.

Минимизация времени получения пользователем ЭД из других ИС. Среднее время τ_i получения ЭД d_j по запросу пользователя ограничивается допустимым временем T_i' т. е

$$\tau_i \le T_i'. \qquad (5)$$

Частота запросов к другим ИС I_i задается в виде

$$Q_i' = \sum_{i=1}^{N}\sum_{j=1}^{m_i} \eta_{ij}' y_{ij}, \qquad (6)$$

где η_{ij}' - частота обращения другим ИС I_i, y_{ij} - индикатор, показывающий наличие требуемый ЭД в других ИС I_i

$$y_{ij} = \begin{cases} 1, & \textit{имеется, если } d_j \in I_i; \\ 0, & \textit{не имеется, если } d_j \notin I_i. \end{cases}$$

Размер памяти, требуемой для хранения ЭД в БД, ограничивается в виде

$$\sum_{i=1}^{N}\sum_{j=1}^{n_i} x_{ij} v_{ij} \le \sum_{i=1}^{n} O_i, \qquad (7)$$

где x_{ij} - количество d_j в H_i, O_i – объем памяти, допустимый для размещения ЭД в БД H_i.

Среднее время представления документа пользователю системой задается в виде

$$\tau_i = \frac{1}{Q_i'}\sum_{i=1}^{N}\sum_{j=1}^{m} \tau_{ij}\eta_{ij}' y_{ij}, \qquad (8)$$

где $\tau_{ij} = \frac{v_{ij}'}{R_i'}$ - время на передачу ЭД d_j с объемом информации v_{ij}' из других ИС I_i, R_i' - пропускная способность каналов переработки информации СЭД.

Подставляя (6) в (8) получаем

$$\frac{1}{Q_i'}\sum_{i=1}^{N}\sum_{j=1}^{m} \tau_{ij}\eta_{ij}' y_{ij} \le T_i' \qquad (9)$$

Стоимость хранения ЭД в БД системы задается в виде

$$S_{xp} = \sum_{i=1}^{N}\sum_{j=1}^{n_i} s_i^{xp} x_{ij} \left(\sum_{k=1}^{n_i} met_{ijk} v_{ijk}' + \sum_{k=1}^{n_i} con_{ijk} v_{ijk}''\right) \qquad (10)$$

где v_{ij} и v_{ij}'' - элементы вводимого и эталонного документов, $s_i^{xp} = \frac{W_i^{xp}}{V_i^{xp}}$ - стоимость хранения единицы информации в БД H_i, W_i^{xp} - стоимость формирования БД H_i,

$$W_i^{xp} = W_i^{во} + W_i^{оо},$$

$W_i^{во}$ - стоимость аренды оборудования, $W_i^{во} = W_i^{o} rd$, W_i^{o} - стоимость оборудования; r - амортизационный коэффициент; d - дисконтные ставки; $W_i^{оо}$ - стоимость обслуживания оборудования;

$$W_i^{xp} = W_i^{зп} + W_i^{mn} + W_i^{лп} + W_i^{a},$$

$W_i^{зп}$ - заработная плата оператора, W_i^{mn} -стоимость технической поддержки, $W_i^{лп}$ - стоимость лицензионной политики, W_i^{a} - стоимость аренды V_i^{xp} - объем хранилище БД H_i.

Стоимость представления ЭД пользователю из БД H_i задается, как

$$s_{nxp} = \sum_{i=1}^{N}\left(\frac{1}{Q_i'}\sum_{i=1}^{n_i} s_i^{nxp}\tau_{ij}\eta_{ij}' x_{ij}\right), \qquad (11)$$

где s_i^{nxp} - стоимость получения единицы информации из БД H_i, $s_i^{nxp} = \left(\frac{W_i^{nxp}}{V_i^{nxp}}\right)R_i$ и $W_i^{nxp} = W_i^{во} + W_i^{m}$, W_i^{m} - тарифная стоимость получения информации из БД H_i, V_i^{nxp} - объем информации, переданной из БД H_i.

Стоимость представления пользователю ЭД из других ИС задается, как

$$s_{пинф} = \sum_{i=1}^{N}\left(\frac{1}{Q_i^{'}}\sum_{j=1}^{m_i} s_i^{пинф}\tau_{ij}\eta_{ij}^{'}x_{ij}\right), \qquad (12)$$

где $s_i^{пинф}$ - стоимость получения единицы информации из других ИС I_i;

$$s_i^{пинф} = \left(\frac{W_i^{пинф}}{V_i^{пинф}}\right)R_i \text{ и } W_i^{пинф} = W_i^{бо} + W_i^{m}.$$

Общая стоимость оптимизации алгоритма повышения достоверности информации на всех этапах переработки документа задается в виде

$$S_{общ} = \lambda_i(S_{хр} + S_{пхр} + S_{пинф}), \qquad (13)$$

где λ_i - коэффициент выигрыша в достоверности информации.

Оптимизация достоверности информации коллекции документов задается в виде

$$S_{опт} = \lambda_1\sum_{i=1}^{N}\sum_{j=1}^{n_i} s_i^{xp}x_{ij}\left(\sum_{k=1}^{n_i}\nu_{ijk}^{'} + \sum_{k=1}^{n_i}\nu_{ijk}^{*}\right) + $$
$$+\lambda_2\sum_{i=1}^{N}\left(\frac{1}{Q_i^{'}}\sum_{i=1}^{n_i} s_i^{пхр}\tau_{ij}\eta_{ij}^{'}x_{ij}\right) + \qquad (14)$$
$$+\lambda_3\sum_{i=1}^{N}\left(\frac{1}{Q_i^{'}}\sum_{j=1}^{m_i} s_i^{пинф}\tau_{ij}\eta_{ij}^{'}x_{ij}\right) \to \min.$$

Вычислительная схема задачи реализована на основе динамического программирования [8].

2.3. Методика оценки выигрышей по критериям эффективности СЭД

Трудоемкость обработки информации документа по каждой операции t_i оценивается, как

$$t_i = \frac{Q_i}{N_i},$$

где Q_i - объем обрабатываемой информации, N_i - информационная емкость документа (нормативный объем информации).

Суммарная трудоемкость обработки информации, связанный с n – операциями по документу определяется в виде

$$T_{общ} = \sum_{i=1}^{n} t_i.$$

Выигрыш в трудоемкости обработки информации документа определяется в виде

$$K_T = 1 - \frac{T_{опт}}{T_{общ}},$$

где $T_{опт}$ - трудоемкость обработки информации, связанный с применением механизмов повышения качества функционирования СЭД.

Аналогично, коэффициент выигрыша в стоимости обработки информации документа задаются в виде

$$K_s = 1 - \frac{S_{опт}}{S_{общ}},$$

где $S_{опт}$ - стоимость обработки информации на основе механизмов повышения качества функционирования СЭД, $S_{общ}$ - стоимость обработки по всей информационной емкости документа.

Для оптимизации качества функционирования СЭД по функционалу задаются следующие ограничения по:

трудоемкости обработки информации $T \le K^{*}$;

стоимости обработки информации $S \le Q^{*}$;

T и S - критерии трудоемкости и стоимости;

Q^{*} и K^{*} - ограничения на значения критериев T и S.

Вводится интервальные ограничения на область допустимых значений критерия T

$$T \le \lambda_1 + \lambda_2 d \text{ либо } T \le a \cdot K^{*} + b,$$

либо критерия S

$$S \le \lambda_1 + \lambda_2 d \text{ либо } S \le a \cdot Q^{*} + b;$$

где $K^{*} = Q^{*} = d$, $a = \lambda_2$ - верхняя граница значений критерия, $b = \lambda_1$ - нижняя граница значений критерия, d - расстояние между точками границ эффективности механизма оптимизации.

В решениях задачи значения a и b регулируется и подбирается эффективная область функционала оптимизации качества функционирования СЭД.

В исследование коэффициента выигрыша в трудоемкости обработки информации заданы следующие значения переменных:

- общий объем информации в одном документе $8 \cdot 10^6$ бит;

- время обработки информации одного документа 0,008 сек.;

- время обработки коллекции из 100 документов - 0,8 сек.

Рассмотрены варианты оптимизации, в которых вместе типовых предельных ограничений на критерии ставится линейное ограничение.

Такой подход позволяет задавать уровень компенсации между противоречивыми факторами и достигать существенное сужение области допустимых решений.

Получены оптимальные точки для коэффициентов K_s и K_T в области допустимых значений при алгоритме жадного поиска глобального экстремума с линейным ограничением.

Установлено, что реализованный алгоритм обеспечивает достижение требуемого значения функционала эффективности при 1600 итерациях по сравнению с алгоритмом поиска с полным перебором всевозможных вариантов, выполняющегося при 65536 итерациях.

Когда применяются механизмы сегментации и параллельных вычислений, то преимущества алгоритмов обработки ЭД возрастают 10 – 15 раз.

2.4. Программный комплекс повышения достоверности информации.

Разработан и реализован программный комплекс повышения достоверности информации, эффективность которого определяется по критериям T_i - трудоемкости либо S_i стоимости обработки информации, проводимые для обнаружения и исправления ошибок в i –й операции.

Результаты исследования оценивается по усредненным значениям критериев T - трудоемкости обработки информации и S - стоимости обработки информации.

Исследование проведено для коллекции из 100 документов по функционалу

$$T_i = \lim_{T\to\infty} T^{-1} \sum_{i=1}^{100} T_i \lambda_i \to \min .$$

Требуется, чтобы высокое значение, которого должно быть близким или равным величине

$$d(T) = \sup_{\lambda_i \in T} (T_i \lambda_i) .$$

И по аналогичному функционалу

$$S_i = \lim_{S\to\infty} S^{-1} \sum_{i=1}^{100} S_i \lambda_i \to \min ,$$

$$d(S) = \sup_{\lambda_i \in S} (S_i \lambda_i) .$$

Для оптимизации функции $d(T)$ либо $d(S)$ в немалых случаях применяется градиентный метод, реализация которого обусловлена со следующими недостатками:

– трудностью вычислительного характера;

– отсутствием возможности явного аналитического представления функции;

– невозможностью получения ее производных характеристик.

Адаптивная стратегия оптимизации функционала, по приближенной оценке, механизм которой предложен вместо механизма определения точного значения градиента функции. В разработанном комплексе синтезирован алгоритм стохастического поиска с частично наблюдаемой Марковской цепи.

Преимуществом этого алгоритма является то, что он выполняется в условиях минимальной априорной информации.

Установлены следующие важные моменты результатов исследования:

– адаптивная оптимизация позволяет получить устойчивое значение стратегии (точки) на множестве T с точностью до третьего знака;

– вероятность выбора значений параметра T_{adan} оценивается в пределах наилучшей и наихудшей стратегии $T_{\max}$ и $T_{\min}$, которые при адаптивной оптимизации поиска оказалась внутренней точкой в множестве значений T.

– разброс между значениями наилучшего и наихудшего стратегий, полученного на основе алгоритма случайного поиска с перебором всех вариантов оказался достаточно большим;

– целевая функция оптимизации становится чувствительной при настройке переменных, что отражается меньшими значениями коэффициента эффективности алгоритма;

– доказано, что трудоемкость обработки информации алгоритма с адаптивным поиском глобального экстремума функционала на основе стохастической модели с частичной Марковской цепью на порядок меньше, чем трудоемкости алгоритма с механизмом поиска со случайным перебором всевозможных вариантов, который дает также менее точные результаты.

В табл.1 приведены результаты достигнутого рейтинга на выходе комплекса, в графах которой отражена эффективность алгоритмов повышения достоверности информации (АПДИ), а традиционной технологии визуального контроля информации (ТВКИ).

Таблица 1. Результаты задачи оптимизации достоверности информации ЭД

Стратегия	Рейтинговая оценка	Выигрыш по рейтингу		Коэффициент эффективности	Коэффициент потери
		АПДИ в %	ТВКИ в %		
$T_{адап}$	37,76	67,49	32,51	0,94	0,322
$T_{\max}$	42,09	68,75	31,25	0,96	0,539
$T_{\min}$	25,40	48,15	51,85	0,68	0,315

III. ЗАКЛЮЧЕНИЕ

Эффективность программного комплекса повышения достоверности информации исследована на следующих примерах СЭД:

- медицинского учреждения (МУ);
- высшего учебного заведения (Самаркандского Гос. университета);
- машиностроительного предприятия Самаркандской области.

Проанализированы значения коэффициентов выигрыша в достоверности, трудоемкости и стоимости обработки информации при различных, примененных механизмах повышения качества функционирования СЭД.

Благодаря применению комплекса достигнуты следующие результаты:

- трудоемкость и стоимость обработки информации коллекции из 100 документов при выполнении функций контроля исполнения ОРД сокращается с 2,4 часа до 1,1 часа по сравнению показателя традиционной технологии визуального контроля и исправления ошибок;
- документы по запросу пользователя представляются 2-3 раза быстрее;
- коэффициенты трудоемкости и стоимости обработки информации уменьшаются 5-7 раз;
- когда применяются алгоритмы повышения достоверности информации на основе использования семантической избыточности и лексикологического синтеза структуры ЭД, тогда значения коэффициентов трудоемкости и стоимости обработки информации уменьшается в среднем 2-3 раза, а достоверность информации повышается до двух порядков.

ЛИТЕРАТУРА

[1] *Matsuo Y and Ishizuka M* Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools 2004 13(1), 157–169.

[2] *Alemeh Matani, Hamid Reza Naji, Hassan Motallebi*, A Fault-Tolerant Workflow Scheduling Algorithm for Grid with Near-Optimal Redundancy, Journal of Grid Computing, 2020. 10.1007/s10723-020-09522-2.

[3] *Wenhao Li*, Hardware Reliability Requirements, Encyclopedia of Big Data Technologies, 2019 10.1007/978-3-319-77525-8, (918-922).

[4] *D.Downey, O. Etzioni, S.Soderland,*: A probabilistic model of redundancy in information extraction. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI'05, Edinburgh, pp. 1034–1041. Morgan Kaufmann, San Francisco (2005)

[5] *G.Hinton, R.aSalakhutdinov*, Discovering binary codes for documents by learning deep generative models Topics in Cognitive Science 2011 Volume 3, Issue 1, Pages 74-91

[6] *Akhatov A., Nazarov F., & Rashidov A.* Mechanisms of information reliability in big data and blockchain technologies" ICISCT 2021, 3-5.11, doi: 10.1109/ICISCT52966.2021.9670052

[7] *Akhatov A., Nazarov F., & Rashidov A.* Increasing data reliability by using big-data parallelization mechanisms. ICISCT 2021: Applications, Trends and Opportunities, 3-5.11.2021, doi: 10.1109/ICISCT 52966.2021.9670387

[8] *Dilek Hakkani-Tür, Kemal Oflazer, & Gökhan Tür.* Statistical Morphological Disambiguation for Agglutinative Languages. In Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000).

[9] *Jumanov I.I. & Karshiev Kh. B.* Increasing the reliability of full text documents based on the use of mechanisms for extraction of statistical and semantic links

of elements, 2020 ICISCT 2020: International Conference on Information Science and Communications Technologies, DOI: 10.1109/ICISCT50599.2020.9351397

[10] *Jumanov I.I. & Karshiev Kh. B.* Mechanisms for optimization of detection and correction of text errors based on combining multilevel morphological analysis with n-gram models, Journal of Physics: Conference Series, DOI 10.1088/1742-6596/1546/1/012082, 2020

[11] *Jumanov I.I. & Karshiev Kh. B.* Optimization of Transmission and Processing of Information of Electronic Documents Based on Parallel Computing Technology, ICTACS 2022: Proceedings of International Conference on Technological Advancements in Computational Sciences, DOI: 10.1109/ICTACS56270.2022.9988150, 2022.

[12] *Kemal Oflazer, Sergei Nirenburg, &Marjorie McSchane.* Bootstrapping Morphological Analyzers by Combining Human Elicitation and Machine Learning. Computational Linguistics, 27- 1: 2001, 59-85.

Поступила в редакцию 28.09.2023

Цитирование: *Жуманов И.И., Каршиев Х.Б.* (2023). Повышение качества функционирования систем электронного документооборота на основе механизмов использования семантической избыточности. *Международный журнал теоретических и прикладных вопросов цифровых технологий*, 4(6), –С. 67-74.

IMPROVING THE QUALITY OF ELECTRONIC DOCUMENT MANAGEMENT SYSTEMS BASED ON MECHANISMS OF USING SEMANTIC REDUNDANCY

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

[1] Samarkand State University named after Sharof Rashidov, Samarkand, Uzbekistan

xusan2005@mail.ru

Abstract. *Methodological foundations have been developed for solving problems of optimizing search, storage, and processing of information according to the criteria of reliability, labor intensity and cost. Estimates of the time and cost of input, transmission, storage, processing, exchange of documents, detection and correction of information errors are proposed based on mechanisms for using semantic redundancy implemented and lexographic synthesis of the document structure. The effectiveness of mechanisms for using statistical, logical, semantic and structural-technological connections of document elements has been studied. A computational scheme for solving an optimization problem has been developed and implemented based on the use of adaptive methods of stochastic random search, truncated Markov chain modeling and dynamic programming. A software package has been implemented to increase the reliability of information based on the use of adaptive random search, segmentation and lexicological synthesis of the structure*.

Keywords: *efficiency, electronic document management system, labor intensity and cost of information processing, information reliability, optimization, optimization strategy, stochastic model.*