SCIENTIFIC JOURNAL OF SAMARKAND UNIVERSITY

Volume 2019 Article 2

6-27-2019

Effectiveness analysis of generalization of algorithms for increasing information reliability based on usage of information redundancy of electronic documents

I.I. Jumanov
Samarkand State University, olimjondi@mail.ru

Kh.B. Karshiyev
Samarkand State University

ILMIY AXBOROTNOMA INFORMATIKA 2019-yil,3-son

УДК: 519.681.5

АНАЛИЗ ЭФФЕКТИВНОСТИ ОБОБЩЕНИЯ АЛГОРИТМОВ ПОВЫШЕНИЯ ДОСТОВЕРНОСТИ ИНФОРМАЦИИ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ ИНФОРМАЦИОННОЙ ИЗБЫТОЧНОСТИ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ

И.И. Жуманов., Х.Б. Каршиев

Самаркандский государственный университет

E-mail: Olimjondi@mail.ru

Аннотация. Предложены подходы, направленных применению естественной избыточности для повышения достоверности информации в виде логических, семантических, технологических, статистических связей, свойств и отношений элементов и ключевых концептов документа. Разработанные методы и алгоритмы основаны на механизмы уточнения, корректировки и контроля значений элементов, признаков, атрибутов, концептов документа. Реализован модифицированный ассоциативный семантический сеть поиска образа - документа путем сегментации в виде его фрагментов и составных частей.

Ключевые слова: электронный документ, достоверность информации, естественная избыточность, логические и семантические связи, отношение, семантический сеть поиска, сегментация, обобщенный алгоритм обработка данных.

Elektron hujjatdagi axborot ortiqchaligidan foydalanish asosida ma’lumotlar ishonchligini oshiruvchi algoritmlarni umumlashtirishning samaradorligi tahlili

Annotatsiya. Hujjat elementi va kalit konseptlari mantiqiy, semantik, texnologik, statistik bog‘lanishi, xossa va munosabatlari ko‘rinishida tabiiy ortiqchalikdan foydalanishga qaratilgan yondoshuvlar taklif etilgan. Ishlab chiqilgan usul va algoritmlar hujjat elementi, belgi, atribut, konseptlari qiymatlarini nazorat, aniqlashtirish, tahrir qilish mexanizmlariga asoslanadi. Hujjat – obrazini fragment va tarkibiy qismlari sifatida segmantatsiyalash hamda qidirishni amalga oshiruvchi takomillashtirilgan assotsiativ semantik tarmoq joriylashtirilgan.

Kalitso‘zlar: elektron hujjat, informatsiya ishonchligi, tabiiy ortiqchalik, mantiqiy vasemantik bog‘lanish, munosabat, qidirish semantik tarmog‘i, segmantlash, ma’lumotga ishlov beruvchi umumlashgan algoritm.

Effectiveness analysis of generalization of algorithms for increasing information reliability based on usage of information redundancy of electronic documents

Abstract. Approaches are proposed to use natural redundancy to increase the reliability of information in the form of logical, semantic, technological, statistical relationships, properties and relations of elements and key concepts of the document. The developed methods and algorithms are based on the mechanisms of refinement, adjustment and control of the values of elements, features, attributes, concepts of the document. Implemented a modified associative semantic network of image search - a document by segmentation in the form of its fragments and components.

Keywords: electronic document, reliability of information, information redundancy, logical and semantic links, relation, semantic search network, segmentation, generalized data processing algorithm.

Актуальность темы.

В системах электронного документооборота (СЭД) предприятий и учреждений циркулируют документы, которые представляются для обработки и хранения в базах данных (БД) электронных документов (ЭД) в различных форматах, в частности в виде офисных, отсканированных бумажных документов, web-страниц, графических изображений, чертежей, видео файлов и т.д[1].

Ключевой проблемой повышения эффективности СЭД является обеспечение достоверности, точности, полноты обработки информации, релевантности документов, значения которых снижаются при переносе данных на машинные носители, передаче по каналам связи, вводе информации, а также по другим субъективным причинам. В условиях СЭД вероятность ошибок в информации, допускаемых человеком –оператором, средствами сканирования и распознавания находится в пределах 10^{-2} - 10^{-3}, вероятность

ошибок, обусловленных из-за помех в каналах связи -10^{-3} -10^{-4}, сбоев и отказов электронного оборудования 10^{-4} - 10^{-5}. Среди них, наиболее уязвимым звеном является ввод информации [6].

В современных инфокоммуникационных сетях широкое распространение нашли программные и аппаратные методы передачи информации, основанных на использовании кодовой избыточности сообщений.

Эффективное выявление и коррекции ошибок информации документов при вводе можно обеспечивать также, такими стандартными средствами, как применением БД, систем управления базами данных (СУБД), базы знаний (БЗ) и экспертных систем [2, 3].

В [5]доказано эффективность применения естественной информационной избыточности, обусловленной из-за логических, семантических, технологических, статистических связей, свойствами и отношениями между элементами и ключевыми концептами документов, инструментарии которых в свою очередь дают возможностей достоверному вводу информации на основе структурно-технологической, статистической, естественной, семантической информационной избыточности.

Методы и алгоритмы повышения достоверности информации, использующие избыточности различной природы позволяет построить, программные средства, отличающиеся своей простотой и дешевизной реализации и возможностью обеспечения высокой достоверности, релевантности, точности, полноты обработки информации с менее трудоёмкими операциями и низкой стоимости.

Основные подходы и принципы повышения достоверности информации ЭД. Предложены подходы, направленные разработке методов и алгоритмов повышения достоверности информации, основывающихся на инструментарии: БД и БЗ; механизмы уточнения, корректировки и контроля достоверности информации элементов, признаков, терминов, атрибутов, концептов документа; ассоциативной семантической сети поиска; сегментации пространства, распознавания и классификации документов в виде фрагментов, составных частей образа.

В СЭД достоверность полнотекстовых ЭД обеспечивается на основе лингвистического, графематического, морфологического, *n*-граммного, семантического анализов, а документы смешенного алфавита контролируются на основе логического контроля, метрик близости, а также по разрешенным интервалам, пороговым границам и другими методами.

Метод графематического анализа направлен выделению элементов из структуры текста, в частности параграфа, абзацев, предложений, отдельных слов и других концептов документов. Методы морфологического и *n*-граммного анализов направлены определению морфологических и *к*-граммных структурных характеристик текста в виде слова, словоформы и др. [5].Метод синтаксического анализа направлен определению синтаксической зависимости слов в предложении, тесной связи между элементами, семантики и синтаксисы контекста [6].

В работах [3-6] эффективность отмеченных методов повышения достоверности информации исследована по критериям достоверности, сложности (трудоемкости) и стоимости обработки информации.

Алгоритм повышения достоверности документа на основе семантической сети поиска. На рис. 1 представлен обобщенный алгоритм повышения достоверности информации.

Разработанная схема повышения достоверности полнотекстового документа базируется на семантический сеть поиска, которая отражается множеством элементов, концептов со связями и отношениями.

Установлено, что эффективность алгоритма зависит от числа поисковых элементов, ключевых концептов, влияют также длина слов в предложение, числа правил и архитектура сети поиска. Например, значение коэффициента трудоемкости обработки информации увеличивается экспоненциально с увеличением количество поисковых элементов и концептов.

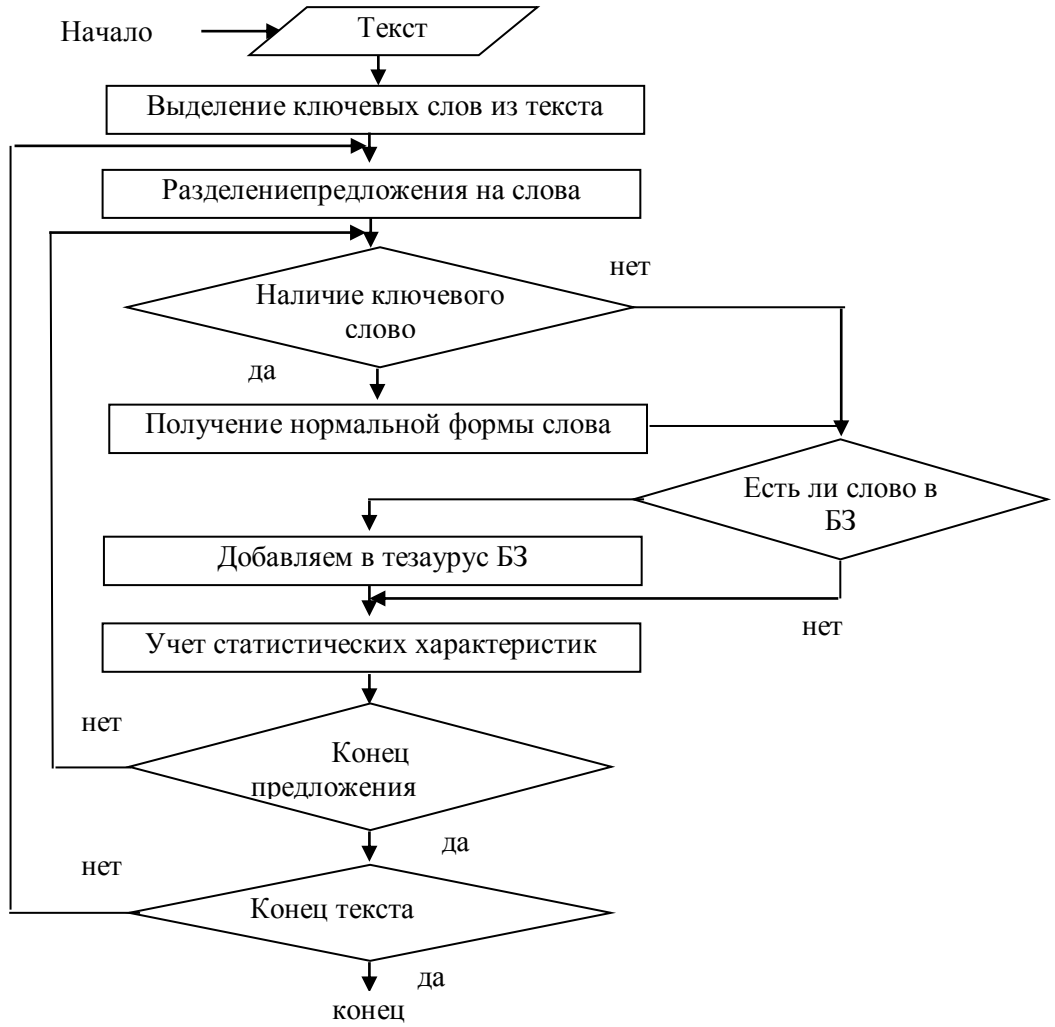

Рис.1. Обобщенный алгоритм повышения достоверности информации.

В табл.1 дана предложенная методика (механизмы, математические модели и выражения расчета, описание обозначений) для построения механизмов и совершенствования обобщенного алгоритма.

Эффективность обобщенного алгоритма повышается за счет применения модели предпочтений, субтрактивных отношений концептов, в результате которых достигается уточнение, проверка достоверности документа, корректировка контекста.

Сравнительный анализ эффективности обобщенного алгоритма. Для сравнительного анализа эффективности обобщенного алгоритма рассмотрен традиционный механизм семантической сети поиска, в котором вычисляется числа вариантов перебора, используемых в дальнейшем в качестве ограничения на область поиска объектов.

Выполнено разбиение (сегментации) пространства элементов и признаков каждого концепта на равные группы, которые затем классифицированы. Благодаря сегментации достигается существенное сокращение число вариантов перебора при поиске документов.

На рис.2 проиллюстрирована зависимость коэффициента трудоёмкости обработки информации по обобщенному алгоритму, включающего механизмом поиска с перебором вариантов, сегментирования, распознавания и классификации, повышения достоверности информации, где по оси ординат измеряется рейтинговое значение показателя эффективности, задаваемого по 50балльной шкале измерения.

Разработан пользовательский интерфейс обобщенного алгоритма обработки информации с семантической сети поиска, которая динамически регулирует навигационную структуру.Каждое действие пользователя инициирует изменение весовых коэффициентов отношений между концептами и модели предпочтений.

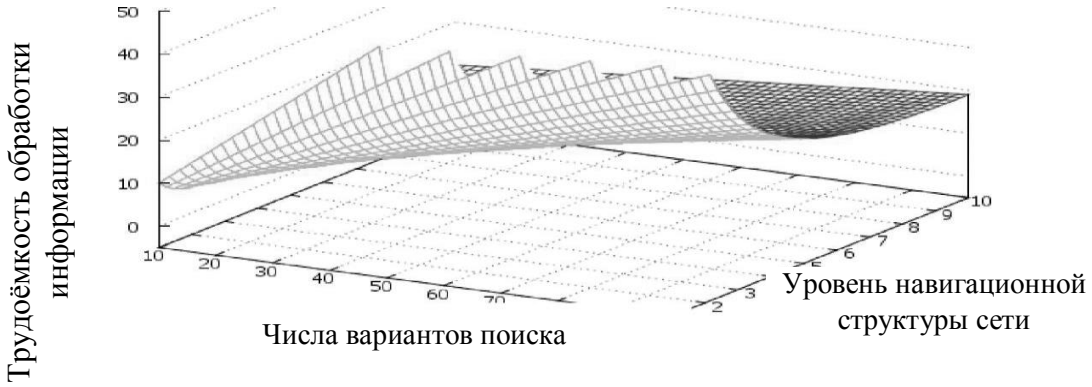

Рис.2. Эффективность обобщенного алгоритма по коэффициенту трудоёмкости обработки информации.

На рис.3 показана схема обобщенного алгоритма со интерфейсом пользователя, который позволяет визуализировать следующие результаты: информационную потребность пользователя; множество концептов поиска в семантической сети; число пересекающихся связей между вершинами; разделения множество концептов на подмножества с последующим их размещением на отдельных плоскостях многомерного интерфейса пользователя; деление множество концептов семантической сети по типам отношений; деление связанных концептов одной плоскости в виде горизонтальных и связанных концепты различных плоскостей в виде вертикальных линии.

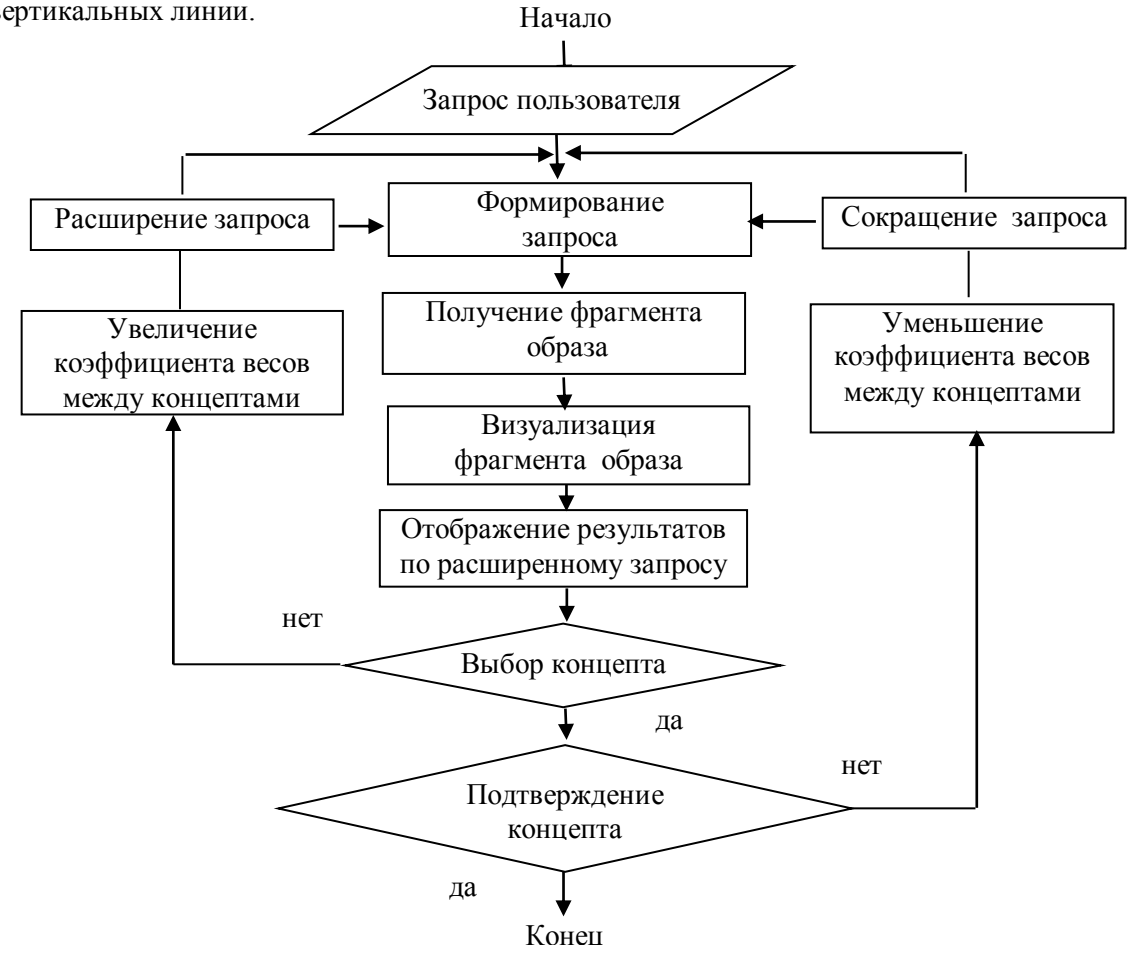

Рис. 3. Схема адаптации пользовательского интерфейса.

Таблица 1

Методика построения механизма для обобщенного алгоритма обработки информации документа

№	Механизм	Модель (математическое выражение) расчета	Описания обозначений
1	2	3	4
1	Определения сходства типов концептов документа	$\forall_{c_i} : E_q(c_i, c_j) = \frac{x / len}{\max(l_i, l_j)}$ $c_i \in C^D, c_j \in C^{KB}$ $i = \overline{1, N_D}, j = \overline{1, N_{KB}}$	C^D -множество концептов в документе D; L^D - множество типов отношений в документе D; C^D_{Syn}, C^{KB}_{Sym} - множество синонимов, соответственно концепта документа и модели KB.
2	Определения сходства структуры концепта документа с контекстом	$\forall_{c_{ш}} : Poseq(c_{i,} c_j) = \frac{\left\|C^D_{Hyp} \cap C^{KB}_{Hyp}\right\|}{\|C^{KB}_{Hyp}\|},$ $c_i \in C^D, c_j \in C^{KB},$	$Poseq()$ - сходства структурной части двух концептов документа; C^D_{Hyp}, C^{KB}_{Hyp} - множество гипонимов концепта c_i u c_j.
3	Задания пороговой функции оценки среднего сходства полноты документа	$f(c_i, c_j) = \frac{a \cdot Eq(c_i, c_j)}{3} + \frac{b \cdot Poseq(c_i, c_j)}{3} + \frac{c \cdot Syn(c_i, c_j)}{3} > z$	z – значение пороговой оценки; a, b, c – значения коэффициентов полноты документа.
4	Ранжирования оценки релевантности документа в соответствии с весовыми коэффициентам и элементов либо концептов документа	$G_i^d = \{g_k \mid (\forall g_k, g_m \exists l \in L :$ $: w(c_k, c_m) > x) \wedge$ $\wedge, l_{Hyp}(c_m, c_z) \in L)\}$ $k, m, z = \overline{1, N_G}, z = \overline{1, N_L},$ $G = \{g \mid \exists c_i = g\}$	G_i^d - i-я группа элементов документа d-го уровня навигационной структуры сети поиска; C - множество концептов в документе; L - множество отношений между концептами документа;
5	Ранжирования с учетом весовых коэффициентов отношений концептов документов	$R(d_k) = \sum_{L_{d_k}} (f\ (\overline{w_k}, r) - \sum_{L'_{d_k}} f\ (\overline{w_k}, r))$ $L_{d_k} = \{l^d \mid (c_i, c_j \in d_k)$	$\overline{w_k}$ - $к$ - я компонента вектора весовых коэффициентов отношений l концептов; x -коэффициент, включения отношений концептов в расширенный запрос;

References

1. Bessonov, S. V. Optimizasiya elektronnogo dokumentooborota v korporativnix sistemax: dis. kand. ekon. nauk. M., 2000 g.187 s.
2. Norenkov I.P., Uvarov M.Yu. Baza i generator obrazovatelnix resursov // Informasionniye texnologii, 2005, № 9, s. 60-65.
3. Lukashevich N. V. Tezaurusi v zadachax informasionnogo poiska. M.: Izd-vo Moskovskogo universiteta, 2011. 512 s.
4. Jumanov I.I Konseptualniye prinsipi i metodi kontrolya dostovernosti informasii v strukture paketov peredachi dannix na osnove statisticheskoy izbitochnosti //«Ilmiy tadqiqotlar axborotnomasi» ilmiy-nazariy, uslubiy jurnal. – Samarqand: SamDU, 2013. - №1 (77) – 39-49 b.
5. Jumanov I.I., Karshiyev X.B. Metodi obespecheniya dostovernosti elektronnix dokumentov na osnove strukturnoy izbitochnosti i leksikologicheskogo sinteza // «Nauka i mir», Mejdunarodniy nauchniy jurnal, Izd-vo «Nauchnoye obozreniye», Volgograd. – №3(55), Tom 1, 2018. – s. 49-51.
6. JumanovI.I., Karshiyev X.B. Expanding the possibilities of instruments to improve the information reliability of electronic documents of industrial management SYSTEMS// Tenth World Conference “Intelligent Systems for Industrial Automation”, WCIS-2018,25-26 October 2018, Tashkent, Uzbekistan, ISBN: 933609-37-2-2018, , b-Quadrat Verlag-86916 Kaufering, –312-316 p.