ТЕХНОЛОГИИ ПОВЫШЕНИЯ ДОСТОВЕРНОСТИ ИНФОРМАЦИИ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ НА ОСНОВЕ *N*-ГРАММНЫХ МОДЕЛЕЙ ЕСТЕСТВЕННОГО ЯЗЫКА

Исроил Ибрагимович ЖУМАНОВ
профессор
доктор технических наук
Самаркандский государственный университет
Самарканд, Узбекистан
husan-qarshiyev@samdu.uz
Хусан Беркинбаевич КАРШИЕВ
доцент
доктор философии (PhD) по техническим наукам
Самаркандский государственный университет
Самарканд, Узбекистан
xusan2005@mail.ru

Аннотация

В статье представлена информация о методах повышения достоверности информации в системах электронного документооборота. Также приведены представления о механизмах вероятности исправления ошибок, кластеризации и оценки достоверности информации с использованием вероятности необнаруженных ошибок.

Ключевые слова: системы электронного документооборота, достоверность информации, информационная избыточность, вероятность необнаруженных ошибок, кластеризация, *n*-грамм, программный комплекс.

ТАБИИЙ ТИЛ *N* - ГРАММЛИ МОДЕЛЛАРИ АСОСИДА ЭЛЕКТРОН ҲУЖЖАТЛАР МАЪЛУМОТЛАРИНИНГ ИШОНЧЛИЛИГИНИ ОШИРИШ ТЕХНОЛОГИЯЛАРИ

Исроил Ибрагимович ЖУМАНОВ
профессор
техника фанлари доктори
Самарқанд давлат университети
Самарқанд, Ўзбекистон
husan-qarshiyev@samdu.uz
Ҳусан Беркинбаевич ҚАРШИЕВ
доцент
техника фанлари бўйича (PhD) фалсафа доктори
Самарқанд давлат университети
Самарқанд, Ўзбекистон
xusan2005@mail.ru

Аннотация

Мақолада электрон ҳужжат айланиш тизимларида ахборот ишончлилигини оширувчи усуллар ҳақидаги маълумотлар ёритилган. Шунингдек, хатолар аниқланмаслиги эҳтимоли ёрдамида ахборот ишончлилигини баҳолаш, кластерлаш ва хатоларни тузатиш эҳтимоли

механизмлари ҳақидаги фикрлар баён этилган.

Таянч сўзлар: электрон ҳужжат айланиш тизимлари, ахборот ишончлилиги, ахборот ортиқчалиги, хатолар аниқланмаслиги эҳтимоли, кластерлаш, *n*-грамм.

В мировой практике информационных технологий в последние годы широко внедряются системы автоматизации управления, в них выделяется востребованность к задачам достоверного сбора, передачи, кодирования информации, формирования документов, а также получению эффективных мобильных и производительных технологий систем электронного документооборота (СЭД) [1-4].

Наибольшая часть искажений информации электронных документов (ЭД) обусловливается ошибками человека-оператора, погрешностями средств сканирования и распознавания, а также орфографическими ошибками пользователей. Существует широкий спектр работ, входящих в круг наиболее развитых исследований, которые направлены на разработку методов контроля орфографических ошибок и реализации графематических, лингвистических, морфологических моделей, главным свойством которых является привлечение словаря словоформ большого размера и механизмов извлечения специфических знаний о применяемом естественном языке документов [5-7].

Предпосылкой решения проблем повышения эффективности СЭД является совершенствование традиционных и создание новых инструментов контроля и коррекции ошибок в текстах, выполняемых существенно меньшей трудоемкостью и стоимостью обработки информации. Исследуемая методология составляет наиболее эффективное направление проектирования инструментов, применение которых обосновано приложениями задач распознавания речи, машинного перевода, информационного поиска, контроля орфографии [8].

Доказана перспективность исследования, направленного на создание методов и алгоритмов повышения достоверности информации ЭД, предназначенных для обнаружения и исправления ошибок различной

кратности на основе *n*-граммных структурированных моделей естественного языка (ЕЯ) [9]. Ключевой считается задача повышения достоверности информации с помощью современных протоколов передачи данных, которая основывается на установлении закономерности распределения ошибок при функционировании любой информационной системы. Однако методы, реализуемые по этому подходу для решения проблемы, еще мало изучены.

Исследованы и смоделированы закономерности распределения ошибок в виде *n*-граммных моделей на основе большого корпуса материалов, используемых в деятельности учреждений государственных услуг, а также разработаны алгоритмы фиксирования искаженных элементов информации (букв, слов) в тексте, механизмы применения типичных инструментов поиска, распознавания, кластеризации и структуризации документов и использования частотных характеристик, а также построенных частотных словарей *n*-грамм небольшого объема [10].

Предысторией для разработки методов обнаружения и исправления ошибок *k*-кратности в текстах ЭД на основе *n*-граммной структурированной языковой модели (ЯМ) являются эксперименты К.Шеннона, результаты которых представляются в виде диграмм. К.Шеннон достоверность рассматриваемой буквы считает зависящей только лишь от предыдущих $n-1$ букв. Однако благодаря наличию естественной избыточности информации наиболее вероятная буква следует за рассматриваемой буквой x_i и такое событие предсказывается на основе предыдущего набора $x_{i-1}, x_{i-2}, \ldots, x_{i-n}$. Определение эффективности методов повышения достоверности информации на основе естественной избыточности связано с исследованием энтропийных характеристик элемента в документе, которые отражаются в виде матрицы с размерностью $n \cdot s$, $n, s \geq 1$. Разработана методика вычисления условной энтропии для набора из ($k+1$) элементов матрицы (k-грамм) [4]. Причем для малых k должно выполняться следующее соотношение

$$H(x_{ij} / x_{ij}^{(1)}, x_{ij}^{(k)}) = H(x_{ij}, \overline{x_{ij}^{(1)}, x_{ij}^{(k)}}) - H(x_{ij}^{(1)}, x_{ij}^{(k)});$$

$$H(\|x_{ij}\|) = H(x_{ij}, \overline{x_{ij}^{(1)}, x_{ij}^{(k)}}) + \sum_{r=k+1}^{ns-1} \min H(x_{ij}^{(r)} / x_{ij}^{(1)}, x_{ij}^{(k)}), \quad (1)$$

которое дает достаточно близкие значения оценки условной энтропии.

Оценки условной энтропии k-грамм находятся, как

$$\tilde{H}(\overline{x_{ij}^{(1)}, x_{ij}^{(k)}}) = -\sum_{\xi_{ij}^{(1)} \xi_{ij}^{(k)}} \tilde{P}(\overline{\xi_{ij}^{(1)}, \xi_{ij}^{(k)}}) log\, \tilde{P}(\overline{\xi_{ij}^{(1)}, \xi_{ij}^{(k)}}), \quad (2)$$

где $\overline{\xi_{ij}^{(1)}, \xi_{ij}^{(k)}} \in \overline{x_{ij}^{(1)}, x_{ij}^{(r)}}$ – конкретная реализация k-грамм;

$\tilde{P}(\overline{\xi_{ij}^{(1)}, \xi_{ij}^{(k)}})$ – оценка вероятности появления k-грамм.

Рассчитаны условные энтропии k-грамм на основе коллекции из 100 документов в виде монограмм, диграмм, триграмм. Определены объемы требуемой естественной избыточности для повышения достоверности информации документов, значения которой колеблются в $0{,}5 \div 0{,}7$.

Вероятность обнаружения ошибок в информации ЭД

Возможное число искажений в последовательности закодированной информации контрольным кодом (КК) зависит от числа кодовых слов, а их оценки представляются в виде функции

$$f_{KK_i} = \sum_{m=1}^{n} f_{KK_i}^{m}, \quad (3)$$

где $f_{KK_i}^{m}$ – число возможных сочетаний кодовых комбинаций, вызванное k-кратными ошибками. В выражении (3) величина $f_{KK_i}^{m}$ выступает в качестве показателя, который характеризует возможности КК в обнаружении ошибочных позиций букв либо символов в тексте. Однако функция f_{KK_i} учитывает только общее число ошибок информации без разделения их на одиночные, двойные и k-кратные ошибки.

Важно отметить, что в функции f_{KK_i} не учитывается удельный вес каждого типа ошибки и частота их появления.

Предложен подход, по которому эффективность методов повышения достоверности информации оценивается по обобщенной величине F_{KK_i}, которая предлагается в качестве критерия – вероятность необнаруженных ошибок

$$F_{KK_i} = \sum_{m=1}^{n} F_{KK_i}^{m} . \tag{4}$$

Здесь $F_{KK_i}^{m}$ – оценки необнаруженных ошибочных элементов документа, вызванные *m*-кратными ошибками.

Общий вид вероятности необнаруженных ошибок задается, как

$$F_{KK_i}^{m} = \sum_{m=1}^{n} f_{KK_i}^{m} P^{m} , \tag{5}$$

где P^{m} – априорная вероятность *m*-кратных ошибок в информации ЭД.

Вероятность одиночной ошибки обозначим через P^{1}, которая предполагается независимым событием. Тогда вероятность двукратных ошибок равна произведению вероятностей одиночных ошибок, как $P^{2} = \left(P^{1}\right)^{2}$. Вероятность трехкратных ошибок оценивается, как $P^{3} = \left(P^{1}\right)^{3}$, а вероятность *m*-кратных ошибок в виде $P^{m} = \left(P^{1}\right)^{m}$.

Исходя из этого выражение (5) преобразуется в виде

$$F_{KK_i} = \sum_{m=1}^{n} f_{KK_i}^{m} \left(P^{1}\right)^{m} . \tag{6}$$

Правомерно полагать, что вероятность возникновения одиночной ошибки в информации документа примерно равна $2 \cdot 10^{-4}$.

Тогда достоверность концепта документа, включающего *t* информационных элементов, оценивается по вероятности $\approx 2 \cdot 10^{-4} t$. Исходя из этих принципов формулу (6) перепишем в виде

$$\begin{aligned} F_{KK_i} &= 2 \cdot 10^{-4} t^{1} f_{KK_i}^{1} + 2^{2} \cdot 10^{-8} t^{2} f_{KK_i}^{2} + 2^{3} \cdot 10^{-12} t^{3} f_{KK_i}^{3} + \\ &+ 2^{4} \cdot 10^{-16} t^{4} f_{KK_i}^{4} + \ldots + 2^{m} \cdot 10^{-4m} t^{m} f_{KK_i}^{m} \end{aligned} \tag{7}$$

или $$F_{KK_i} = \sum_{m=1}^{n} \left(2 \cdot 10^{-4}\right)^m t^m f_{KK_i}^m \ .$$

(8)

Анализ формулы (7) показывает, что определение и использование в дальнейших расчетах числа ошибочных позиций, вызванных больше чем трехкратными ошибками, не имеет смысла, так как вероятность их появления настолько мала, что практически сводит к нулю все другие члены, начиная с четвертого. В результате чего достоверность элемента документа оценивается в виде

$$F_{кк_l} = \sum_{m=1}^{3} (2 \cdot 10^{-4})^m t^m f_{kk_l}^m . \qquad (9)$$

Далее, подставив в формулу (9) значения величины $f_{кк_i}^m$, получим оценку достоверности информации концепта документа. Причем значение показателя $f_{кк_i}^m$ может быть определено как аналитическим путем, так и на основе экспериментальных исследований. Чем меньше будет значение величины $F_{кк_l}$, тем эффективнее считается метод повышения достоверности информации ЭД.

Реализация изложенных теоретических положений направлена на использование естественной избыточности информации, что обуславливает разработку механизмов выявления искаженных элементов (букв, слов) в тексте ЭД и использования типичных инструментов поиска, распознавания, кластеризации, структуризации, а также определения частотных характеристик, законов распределения *n*-грамм в структуре контекста и механизма контроля и коррекции орфографических ошибок.

Ниже предлагается методика вероятностной оценки объема естественной избыточности, необходимого для эффективной организации алгоритмов повышения достоверности информации ЭД.

Оценки объема избыточности информации ЭД

Решение задачи связано с оценкой условной энтропии одного элемента документа S_i, которая для применения задается средней энтропией по алфавиту при условии, что известен предыдущий элемент $S_k^{/}$

$$H(S_i/S_k^{/}) = -\sum_{k=1}^{2^m} P(S_k^{/})\sum_{i=1}^{2^m} P(S_i/S_k^{/})\log P(S_i/S_k^{/}). \qquad (10)$$

Для вычисления энтропии (10) использованы результаты экспериментальных исследований по определению монограмм, диграмм, триграмм и других многомерных вероятностей вида $P(S_i/S_k^{/})$.

Дана верхняя граница оценки условной энтропии, которая учитывает как неравномерность распределения, так и корреляционные связи между элементами документа. Оценка среднего количества информации, содержащегося в одном элементе концепта документа, задается в виде

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

$$+\sum_{k=1}^{2^m} P(S_k^{/})\sum_{i=1}^{2^m} P(S_i/S_k^{/})\log P(S_i/S_k^{/}) \qquad (11)$$

Оценки количества информации, пригодного для любой перекодировки элементов документа, получены в виде

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

$$\times\log P_H(S) - \frac{P_H(S)}{1-P_0(S)}\log[1-P_0(S)] - \frac{P_H(S)}{1-P_0(S)}\log(2^m-1) = m - \frac{P_H(S)}{1-P_0(S)}\times \qquad (12)$$

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

Отметим характерные моменты результатов:

- при вероятности $P_H(S)=0$ имеем $J(S_i/S_k^{/}) = m$, поскольку $Q(S) = 1 - P_0(S)$;

- при другом граничном условии имеем $Q(S)=0$, так как $P_H(S) = 1 - P_0(S)$ и количество информации $J(S_i/S_k^{/}) = m - \log_2(2^m - 1)$;

- выражение $J(S_i/S_k^{/})$ определяет тот минимум дополнительной информации, который необходим для устранения потери из-за ошибок;

- для кода, когда $n = 7$, $k = 3$, достоверность информации оценивается с вероятностью $Q(S) = (1-P)^m$, $m = 10$.

Вероятность обнаружения ошибок в информации находится в виде

$$P_0(S) = (1-P)^7 + 7P^4(1-P)^3. \qquad (13)$$

Вероятность необнаруженных ошибок в информации оценивается, как

$$P_H(S) = (2^m - 1)P^4(1-P)^3. \qquad (14)$$

Количество информации, требуемое при повышении достоверности элементов документа, определяется в виде:

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

(15)

Учитывая, что $P \ll 1$, данное выражение можно переписать в виде:

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

(16)

Подставив в формулу (16) оценки максимальной и условной энтропии, определим требуемый объем избыточности информации в виде

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

(17)

Механизм обнаружения ошибок в текстах ЭД

Первым принципом решения задачи является разработка механизмов для обнаружения искаженных элементов концепта (в тексте) документа за счет использования словаря словоформ опорного ЕЯ, типичных инструментов поиска, распознавания, кластеризации слов и механизма сравнения контрольного слова со словоформой в словаре. Для повышения достоверности информации предполагается, что образы – буквы или слово уже получены. Устанавливается, имеется ли этот образ в словаре. Если не

установлено его наличие, то считается, что такой образ в словаре словоформ отсутствует. Далее определяются наиболее близкие к нему слова. Реализация такого алгоритма связана с определением рационального объема необходимого для частотных словарей *n*-грамм. Вследствие чего существенно снижается время поиска слова по сравнению с работой алгоритма поиска в чрезмерно большом объеме словаря словоформ, который замедляет контроль и коррекцию ошибок в информации документа. Причем определено, что в любом ЕЯ первая тысяча самых частотных слов покрывает от 70 до 90% всего исследуемого контекста.

Для ускорения процесса поиска, распознавания, классификации реализован механизм упорядочения слов в частотных словарях по частотности встречаемых слов. Наиболее частотные слова выделяются в отдельный класс. Проверка слова на наличие в этом списке даже на начальном этапе алгоритма заметно ускоряет процессы.

Второй принцип решения задачи предложен для совершенствования распространенной модели морфологического анализа путем дополнения механизма применения частотного словаря *n*-грамм.

Исследован вопрос моделирования цепью Маркова. Реализация такого подхода связана с применением механизма разделения фиксированного корпуса текста на заранее заданное количество псевдограмматических классов по условной диграммной вероятности следования слов друг за другом. В рассмотренном подходе применен также механизм использования трех-, четырех- и так далее до *n*-граммных последовательностей.

Определено, что применение *n*-граммных последовательностей более высокого ранга по контексту материала требует большего объема вычислений при поиске, распознавании и классификации. Диграммные, трехграммные структурные модели нередко становятся более предпочтительными, чем модели высшего ранга. Применения 9-граммных последовательностей немного, что добавило к оценке достоверности информации по сравнению с трехграммной моделью.

Для разделения контекста материалов на непересекающиеся классы и кластеризации слов предложен механизм, направленный на односторонний просмотр строки текста слева или же справа. В случае применения двухсторонней модели просмотр строки текста производится поочередно и слева и справа.

Выявлено, что механизм кластеризации слова на основе односторонней модели позволяет значительно быстрее, без заметных потерь обеспечить выделение слов и их разбивку на классы.

Механизм коррекции обнаруженных ошибок информации ЭД

Общую схему механизма поиска для коррекции искаженных элементов (слова) документа можно описать в виде следующих шагов:

– первое слово концепта документа (предложения) ищется в классе наиболее частотных слов;

– если оно там не находится, то поиск осуществляется в классе (классах) слов, которые могут стоять на первом месте предложений;

– если оно и здесь не находится, тогда либо выдается отказ, либо производится поиск по полному словарю словоформ;

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

Алгоритм базируется на построении псевдограмматических классов слов, основанных на *n*-грамм в контексте материалов. Для этого представлен достаточно обширный корпус грамматически правильных узбекских слов пятитомного популярного литературного текста и корпус слов политической литературы 16 томов.

В экспериментальных исследованиях воспользовались корпусом примерно в 0,5 млн слов, в котором имелось около 65 тысяч разных слов. Эти слова с помощью программы, написанной на языке Java, разбиты на 20, 50, 100, 200, 500 классов, полученные результаты использованы при распознавании элементов заданного тестового текста.

Определено, что частота *n*-го слова в таком списке окажется приблизительно *обратно пропорциональной* его порядковому номеру *n*, т.е. рангу этого слова, т.е. первое по используемости слово встречается примерно в два раза чаще, чем второе, и в три раза чаще, чем третье, и т.д.

Разработана и реализована методика применения закона Джорджа Ципфа, которая позволила моделировать вероятностные процессы распределения частоты появления слов в исследуемом контексте материалов.

Исследования проведены по пяти корпусам популярного литературного текста на узбекском языке с общим объемом 404637 слов. Из них первый корпус содержит 79893 слова; второй корпус – 96157 слов; третий – 97369 слов; четвертый – 64841 слово; пятый – 66377 слов. Установлено, что наиболее часто встречается слово «бир» – 5488 раз, что составляет почти 14% от всех появлений слов. Слово «сиз» насчитывает чуть больше, чем 12 %, т.е. 4963 раза; слово «бу» встречается 4191 раз. Необходимо и достаточно 135 пунктов словаря словоформ, чтобы учесть половину корпуса, что доказывает достаточность сформированного объема словаря словоформ. Сравнены усредненные частотные характеристики *k*-грамм в испытанных корпусах текста согласно закону распределения Ципфа при гарантийной вероятности $\Re = 0{,}95$. Сформировано 10 наборов элементов документа, каждый из которых составляет 10^5 десят. знак. В 7 случаях наблюдается подтверждение гипотезы о распределении Ципфа.

Для применения в СЭД учреждений по оказанию государственных услуг населению предложен программный комплекс обнаружения и исправления *k*-кратных ошибок информации документов, который реализован и протестирован в среде фреймворк Сфинкс-4.

Комплекс состоит из различных программных модулей. Модуль «интерфейс» поддерживается MFCC, PLP и LPC механизмами извлечения статистических, динамических и специфических характеристик элементов и свойства документов. Модуль «лингвист» поддерживается различными языковыми моделями, включая CFGs, FSTs и *n*-граммы. Модуль «декодер»

поддерживается процедурами модуля «поиска», включая традиционные инструменты Viterbi или Bushderby, а также механизм параллельного поиска. Модуль «конфигуратор» поддерживается различными модулями и механизмами оптимизации обработки информации.

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА:

1. Келдыш Н.В. Методические основы автоматизированного решения задач ведомственного электронного документооборота. // Науч.метод. сборник № 56. - М., 2009. - С.110-117.
2. Бессонов С.В. Оптимизация электронного документооборота в корпоративных системах: дис. канд. экон. наук. М., 2000 г.,187 с.
3. Гудов А.М. Об одной модели оптимизации документопотоков, реализуемой при создании системы электронного документооборота, сборник «Вычислительные технологии», 2006. - С. 53 - 65.
4. Жуманов И.И., Ахатов А.Р. Оценка эффективности программного комплекса контроля достоверности текстовой информации систем электронного документооборота // «Химическая технология. Контроль и управление» - ТГТУ, Ташкент, 2009- № 2, с. 46-52.
5. J. K. Baker, “The Dragon system - an overview,” in *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 23, no. 1, Feb. 1975, pp. 24–29.
6. G. D. Forney, “The Viterbi algorithm,” *Proceedings of The IEEE*, vol. 61, no. 3, pp. 268–278, 1973.
7. Brown P., Pietra V. D., DeSouza P., Lai J. and Mercer R. Class-based n-gram models of natural language. Computational Linguistics, 1992.18:467–479.
8. Ney H. and Kneser R. Improved clustering techniques for class-based statistical language modelling. In European Conference on Speech Communication and Technology (Eurospeech), 1993. pages 973–976, Berlin.
9. George K. Zipf, The Psychobiology of Language, Houghton-Mifflin, New York, NY, 1935 (http://citeseer.ist.psu.edu/context/64879/0).

10. Akhatov A.R., Jumanov I.I. Improvement of text information processing quality in documents processing systems // 2nd IEEE/IFIP International Conference In Central Asia On Internet ICI-2006, September 19-21, International Hotel Tashkent, Uzbekistan.