

УДК 681.51:57

**МЕТОД СНИЖЕНИЯ ТРУДОЕМКОСТИ КОРРЕКЦИИ ОШИБОК
ПОЛЬЗОВАТЕЛЯ НА ОСНОВЕ СОВМЕСТНОГО ПРИМЕНЕНИЯ
ПОМЕХОЗАЩИТНЫХ КОДОВ И ВИРТУАЛЬНОГО СЛОВАРЯ
ДОПУСТИМЫХ СЛОВ**

В.А. Литвинов, С.Я. Майстренко, Ю.Г. Пилипенко

Институт проблем математических машин и систем НАН Украины

e-mail: takam@post.com

1. Значительная часть первичной информации СППР различного назначения представляет собой классификационные коды, отражающие определенные свойства объектов (коды классификаторов образцов техники, номера банковских счетов, коды налоговых классификаторов и т.п.). Такие коды могут содержать избыточные символы, вычисляемые по определенным алгоритмам теории помехозащитного кодирования и служащие для автоматического исправления ошибок, допущенных на этапах формирования и/или ввода документов. Обзор основных алгоритмов помехозащитного кодирования (ПЗК) в приложении к типовым ошибкам пользователя приведен в [1]. К сожалению, все имеющие практический смысл (т.е., требующие реально приемлемой избыточности) коды ПЗК ограничиваются исправлением ошибок типа замены одного символа (однократных транскрипций) и, как минимум, дополнительно перестановок соседних символов (смежных транспозиций). В табл.1 приведены примеры применения основных кодов.

Предлагаемый метод, позволяющий расширить множество корректируемых ошибок и тем самым снизить трудоемкость корректировки ошибок, заключается в совместном применении кодов ПЗК и анализа обратных искажений («вариаций») ошибочного слова.

2. Примем следующие основные обозначения.

$E_{\kappa}^{(1)}$, $E_{\kappa}^{(2)}$, $E_{нк}$ - соответственно непересекающиеся подмножества ошибок, автоматически корректируемых кодом ПЗК (первично корректируемых), принятых для коррекции методом обратных искажений (вторично корректируемых) и некорректируемых ошибок. $E_{\kappa}^{(1)} \cup E_{\kappa}^{(2)} \cup E_{нк} = E\{E_i\}$.

p_i - относительное количество ошибок класса E_i ;

q, n, n_k - соответственно алфавит, полное количество символов слова, количество контрольных символов.

T - виртуальный словарь, - множество слов, удовлетворяющих условию отсутствия ошибки, определенному принятым алгоритмом ПЗК.

$r = \frac{q^{n_k}}{q^n}$ - коэффициент избыточности кода, определяющий относительное количество

всевозможных комбинаций значений n символов, не удовлетворяющих условию вхождения в T .

3. Общая схема реализации метода включает 2 этапа.

1. Декодирование слова (кода) штатной процедурой используемого кода ПЗК. Если класс обнаруженной ошибки принадлежит $E_{\kappa}^{(1)}$, она автоматически исправляется кодом. Иначе, - переход к этапу 2.

2. Генерация V вариаций в классах ошибок, принадлежащих $E_{\kappa}^{(2)}$ и проверка каждой вариации на соответствие T (т.е. многократное повторение этапа 1 для генерируемых вариаций). В зависимости от результатов проверки возможно автоматическое или

полуавтоматическое (с подтверждение пользователя) исправление в соответствии с алгоритмами, описанными в [2].

В целом, возможны следующие финальные исходы:

- ошибка правильно идентифицирована и исправлена автоматически кодом или по виртуальному словарю (вероятность P_{AK} , трудоемкость F_{AK});
- ошибка идентифицирована и исправлена по виртуальному словарю полуавтоматически (вероятность $P_{ПАК}$, трудоемкость $F_{ПАК}$);
- идентифицировать ошибку не удастся, и она исправляется полностью вручную (вероятность P_{PK} , трудоемкость F_{PK});
- ошибка идентифицирована неверно и автоматически исправлена ложно (вероятность $P_{ЛК}$, трудоемкость $F_{ЛК}$);
- ошибка не обнаружена (вероятность $P_{НО}$).

Примем допущение, что значения вариаций ошибочного слова случайно распределены среди q^n всевозможных слов. Это допущение позволяет при анализе вероятностей исходов воспользоваться соотношениями общей модели испытаний Бернулли.

4. Сравнительный анализ трудоемкости операций коррекции проведем на примере использования алгоритма полуавтоматической коррекции по словарю, который заключается в предъявлении пользователю до m синонимов (вариантов коррекции). При этом, если реальное количество синонимов $z \leq m$, пользователь подтверждает нужный вариант; если $z > m$, то ему сразу предлагается корректировать вручную.

Опущенный нами анализ логико-вероятностных графов, описывающих структуры вероятностей частных исходов [2], для рассматриваемых случаев дает следующие выражения.

Вариант А (использование только кода)

$$P_{AK}^{(A)} = P_1; P_{PK}^{(A)} = (1-r) \cdot (1-\beta) \cdot (1-P_1); P_{ЛК}^{(A)} = (1-P_1) \cdot (1-r) \cdot \beta; P_{НО}^{(A)} = (1-P_1) \cdot r \quad (1)$$

Вариант В (код + анализ синонимов по виртуальному словарю)

$$\left. \begin{aligned} P_{AK}^{(B)} &= P_1; \quad P_{ПАК}^{(B)} = (1-P_1) \cdot (1-r) \cdot (1-\beta) \cdot P_2 \pi(m); \\ P_{PK}^{(B)} &= (1-P_1) \cdot (1-r) \cdot (1-\beta) \cdot P_2 (1-\pi(m)) + (1-P_1) \cdot (1-r) \cdot (1-\beta) \cdot (1-P_2) = P_{PK}^{(1)} + P_{PK}^{(2)}; \\ P_{ЛК}^{(B)} &= (1-P_1) \cdot (1-r) \cdot \beta. \end{aligned} \right\} \quad (2)$$

Здесь $P_1 = \sum_{E_k^{(1)}} p_i$, $P_2 = \frac{\sum_{E_k^{(2)}} p_i}{1-P_1}$, через β обозначено относительное количество

ошибок, ложно корректируемых кодом, а $\pi(m)$ равно [2]:

$$\pi(m) = \sum_{g=0}^{m-1} C_{V-1}^g \cdot r^g \cdot (1-r)^{V-g-1} + \sum_{g=m}^{V-1} \frac{m}{g+1} C_{V-1}^g \cdot r^g \cdot (1-r)^{V-g-1}.$$

При анализе трудоемкостей F , оцениваемых через среднее количество синонимов, обрабатываемых пользователем, примем во внимание следующие соображения и явные соотношения.

- $F_{AK}^{(A)} = F_{AK}^{(B)} = F_{LK}^{(A)} = F_{LK}^{(B)} = 0$.
- $F_{PK}^{(A)} = \gamma \cdot P_{PK}^{(A)}$, где γ - среднее количество символов ошибочного слова, обрабатываемых пользователем при его ручной корректировке.
- Ручная коррекция в варианте В возможна в двух случаях:
 - ошибка $\in E_k^{(2)}$, но не «попала» в m предъявляемых синонимов. Вероятность такого случая обозначена в [2] через $P_{PK}^{(1)}$, а трудоемкость $F_{PK}^{(1)} = \gamma \cdot P_{PK}^{(1)}$;
 - ошибка $\notin E_k^{(2)}$ (вероятность $P_{PK}^{(2)}$). В этом случае необходимо отвергнуть $g = 0, 1, \dots, m$ случайных ложных синонимов и исправить ошибку вручную.

В целом

$$F_{PK}^{(B)} = \gamma \cdot P_{PK}^{(B)} + P_{PK}^{(2)} \sum_{g=0}^m g \cdot P(r, g, V - g) \quad (3)$$

- Полуавтоматическая коррекция в варианте В возможна в двух случаях:
 - ошибка $\in E_k^{(2)}$, попала в m предъявляемых синонимов, при этом количество случайных синонимов $g \leq m - 1$;
 - ошибка $\in E_k^{(2)}$, попала в m предъявляемых синонимов, но $g > m - 1$.

В целом

$$F_{ПАК}^{(B)} = (1 - P_1) \cdot (1 - r) \cdot (1 - \beta) \cdot P_2 \left[\sum_{g=0}^{m-1} \frac{g+2}{2} \cdot P(g, r, V - g - 1) + \frac{m+1}{2} \sum_{g=m}^{V-1} \frac{m}{g+1} \cdot P(g, r, V - g - 1) \right] \quad (4)$$

При снятии ограничений на значение m выражения (3), (4) существенно упрощаются. Так, можно показать, что для $m = z$, где z - полное количество синонимов (или, иначе говоря, для $m \rightarrow V$)

$$F_{PK}^{(B)}(m = z) = P_{PK}^{(B)}(rV + \gamma);$$

$$F_{ПАК}^{(B)}(m = z) = P_{ПАК}^{(B)} \left[1 + \frac{1}{2} r(V - 1) \right].$$

5. Рассмотрим результаты приложения полученных соотношений к весовому коду Сетхи-Раджарамана-Кеньяла [3], - одному из наиболее подходящих для ПЗК алфавитно-цифровой информации.

Описываемый метод является в определенном смысле развитием весовых систем избыточного кодирования реквизитов, используемых для контроля достоверности. Метод требует применения двух избыточных разрядов и обеспечивает одновременное исправление однократных транскрипций и одиночных транспозиций соседних символов. В таблицах 1, 2 приведены иллюстративные расчетные данные для вариантов А и В.

$q = 10, q = 37$

Табл. 1

r	$P_{AK}^{(A)}$	$P_{PK}^{(A)}$	$P_{LK}^{(A)}$	$F_{PK}^{(A)}$	$F^A = F_{PK}$
0,001	6,2210 E-1	3,7747E-1	3,7747E-4	1,8855E+0	1,8855E+0
0,0001	6,2210 E-1	3,7778E-1	3,7778E-4	1,8890E+0	1,8890E+0

Табл.2

q	r	m	$P_{\text{ПАК}}$	$P_{\text{РК}}$	$P_{\text{ЛК}}$	$F_{\text{ПАК}}$	$F_{\text{РК}}$	$F_{\text{ПАК}} + F_{\text{РК}}$
10	0,001	2	2,7621E-1	1,0093E-1	3,7752E-4	2,9077E-1	5,1666E-1	8,0743E-1
		5	2,7682E-1	1,0032E-1	3,7752E-4	2,9255E-1	5,1368E-1	8,0623E-1
		10	2,7682E-1	1,0032E-1	3,7752E-4	2,9255E-1	5,1368E-1	8,0623E-1
	0,0001	2	2,7707E-1	1,0076E-1	3,7786E-5	2,7863E-1	5,0500E-1	7,8363E-1
		5	2,7707E-1	1,0075E-1	3,7786E-5	2,7865E-1	5,0497E-1	7,8362E-1
		10	2,7707E-1	1,0075E-1	3,7786E-5	2,7865E-1	5,0497E-1	7,8362E-1
37	0,001	2	2,7032E-1	1,0682E-1	3,7752E-4	3,1356E-1	5,7331E-1	8,8687E-1
		5	2,7682E-1	1,0033E-1	3,7752E-4	3,3352E-1	5,4357E-1	8,7709E-1
		10	2,7682E-1	1,0032E-1	3,7752E-4	3,3354E-1	5,4356E-1	8,7710E-1
	0,0001	2	2,7699E-1	1,0083E-1	3,7786E-5	2,8253E-1	5,0835E-1	7,9088E-1
		5	2,7707E-1	1,0075E-1	3,7786E-5	2,8276E-1	5,0796E-1	7,9073E-1
		10	2,7707E-1	1,0075E-1	3,7786E-5	2,8276E-1	5,0796E-1	7,9073E-1

В расчетах использованы литературные данные [4] для укрупненного перечня типовых ошибок и вероятностей их появления (E_1 - однократные транскрипции, $p_1 = 0,5557$; E_2 - одиночные вставки символов, $p_2 = 0,1567$; E_3 - одиночные выпадения символов, $p_3 = 0,1204$; E_4 - транспозиции соседних символов, $p_4 = 0,0664$; E_5 - двукратные транспозиции, $p_5 = 0,0322$; E_6 - прочие ошибки, $p = 0,0686$).

Для параметров n, β, γ приняты значения $n = 10, \beta \approx r, \gamma \approx \frac{n}{2} = 5$. Для выбранного кода

$$E_{\kappa}^{(1)} = E_1 + E_4; \text{ для ансамбля корректируемых по словарю ошибок выбрано } E_{\kappa}^{(2)} = E_2 + E_3.$$

Как видно из данных табл.1, 2 дополнение штатной процедуры выбранного алгоритма ПЗК анализом вариаций ошибочного слова по виртуальному словарю допустимых слов позволяет значительно снизить трудоемкость идентификации и коррекции ошибок (с 1,8855 для варианта А до 0,8062÷0,7836 для варианта В). При этом выбираемое значение m имеет оптимум, который для конкретных данных может быть определен численным решением (2), (3), (4).

Как показывают дополнительные расчеты, выбор $E_{\kappa}^{(2)} = E_2 + E_3 + E_5$ дает лучшие результаты, чем $E_{\kappa}^{(2)} = E_2 + E_3$ только для $q = 10$ и значений $r = 0,0001$ и менее.

Список литературы

1. Бояринов И.М., Давыдов А.А. Мамедли Э.М., Смеркис Ю.Б. Использование помехоустойчивого кодирования для защиты информации от ошибок оператора //АТ.-1983.-№2.-С.5-48.
2. Кузьменко Г.С., Литвинов В.А., Майстренко С.Я., Хомак В.І. Алгоритми і моделі автоматичної ідентифікації та корекції типових помилок користувача на основі природної надмірності // Математичні машини і системи.-2004.- №2.-С.134-148.
3. Sethi A.S., Rajaraman V., Kenjale P.S. An error-correcting coding scheme for alphanumeric data // Information Processing Letters.-1988.-Vol.7.-P.72-77.
4. Литвинов В.А., Крамаренко В.В. Контроль достоверности и восстановления информации в человеко-машинных системах.-Киев: Техніка, 1986.-200 с.