УДК 681.51:57

АВТОМАТИЧЕСКОЕ ИСПРАВЛЕНИЕ ОШИБОК ПОЛЬЗОВАТЕЛЯ В ЧЕЛОВЕКО-МАШИННЫХ СИСТЕМАХ. МЕТОДЫ И ХАРАКТЕРИСТИКИ

В.А. Литвинов, С.Я. Майстренко
Институт проблем математических машин и систем НАН Украины
e-mail:litvi@dr.com

1. В настоящее время основным средством массового ввода в ЭВМ текстовой информации большого объема (книги, газеты, журналы и т.п.) является сканер в совокупности с OCR-технологиями распознавания символьной информации и программными инструментами автоматической пост-обработки типа After Scan [1]. Однако, преимущества OCR-технологий, принципиально связанные с большой скоростью ввода и меньшим участием человека в процессе ввода, в значительной мере теряют свое значение при:

- вводе в БД фактографической информации, требующей высокой достоверности;
- вводе в БД информации с документов сложной структуры и/или плохого полиграфического качества, рукописных документов;
- отсутствии оформленных первичных документов, - например, вводе коротких сообщений в режиме диалога.

Во всех этих случаях клавиатура пока сохраняет свое значение как основной инструмент ввода данных в ЭВМ. Поэтому задачи повышения эффективности традиционного общения пользователь-ЭВМ с помощью клавиатуры по-прежнему остаются актуальными.

2. Одним из общих направлений совершенствования технологий ввода с клавиатуры является автоматическое обнаружение и исправление ошибок пользователя. Среди ряда известных методов автоматического исправления ошибок, (в частности, “кодовых” методов [2]), особое место занимают словарные методы, основанные на анализе словаря-справочника допустимых значений. Эти методы не требуют, в отличие от кодовых, введения специальной избыточности и позволяют обнаруживать и исправлять все типовые ошибки пользователя.

Сущность общего метода исправления ошибок по словарю заключается в генерации V обратных искажений ("вариаций" [3]) ошибочного слова различными типовыми ошибками пользователя (в частности, однократными транскрипциями E_1, вставками E_2, пропусками E_3, смежными транспозициями E_4, двукратными транскрипциями E_5) и проверки допустимости значения вариации по словарю – справочнику объемом N $n-$символьных слов, представленных в алфавите q. В зависимости от решений, принимаемых в результате проверки, возможны 4 основных алгоритма:

- АКМВ – автоматическая корректировка по вариации, соответствующей максимальной вероятности ошибки;
- АОК – автоматическая однозначная корректировка по единственной совпавшей вариации;
- ПАКМВ – полуавтоматическая корректировка с участием пользователя по вариации, соответствующей максимальной вероятности ошибки;
- АПАКК – автоматическая/полуавтоматическая комбинированная корректировка.

3. Вероятностные характеристики алгоритмов определяются значениями вероятностей следующих исходов, составляющих полную группу возможных сообщений:

- ошибка исправлена правильно автоматически – вероятность P_{AK};
- ошибка исправлена правильно полуавтоматически за m попыток (или выбором из m альтернатив) – вероятность $P_{ПАК}^{(m)}$;

- ошибка исправлена автоматически неправильно (ложная корректировка) – вероятность $P_{ЛК}$;
- ошибка исправлена "вручную" пользователем – вероятность $P_{РК}$.

В основу оценки вероятностных характеристик положена модель испытаний Бернулли, определяющая вероятность $P(l)$ в точности l случайных совпадений при проверке V вариаций:

$$P(l) = C_V^l \left(\frac{N}{q^n}\right)^l \left(1-\frac{N}{q^n}\right)^{V-l}.$$

В табл. 1 приведены сводные вероятностные характеристики, рассчитанные для $N = 10000, q = 10$ и ансамбля ошибок $E_{1,2,3,4}$; значения группового параметра α приведены в табл.2, где $r = N/q^n$.

Табл.1

Алгоритм	α	$P_{АК}$	$P_{ПАК}$	$P_{РК}$	$P_{ЛК}$
A	1	4,8962E-1	-	2,5695E-2	4,7468E-1
	2	8,9124E-1	-	9,9022E-2	9,6346E-3
	3	8,9910E-1	-	1,0078E-1	1,2008E-4
	4	8,9920E-1	-	1,0080E-1	1,4320E-6
B	1	2,3153E-1	-	7,2378E-1	3,4688E-2
	2	8,8342E-1	-	1,1472E-1	1,7527E-3
	3	8,9900E-1	-	1,0097E-1	2,2070E-5
	4	8,9920E-1	-	1,0080E-1	2,6309E-7
C	1	-	7,4772E-1	2,4228E-1	≈ 0
	2	-	8,9906E-1	1,0084E-1	≈ 0
	3	-	8,9920E-1	1,0080E-1	≈ 0
	4	-	8,9920E-1	1,0080E-1	≈ 0
D	1	2,3153E-1	5,5325E-1	1,7053E-1	3,4688E-2
	2	8,8342E-1	1,5686E-2	9,9038E-2	1,7527E-3
	3	8,9900E-1	1,9600E-4	1,0078E-1	2,2070E-5
	4	8,9920E-1	2,3379E-6	1,0080E-1	2,6309E-7

Табл.2

α	$n_{ср}$	r	V
1	6	10^{-2}	135
2	8	10^{-4}	177
3	10	10^{-6}	219
4	12	10^{-8}	261

Громоздкие, но точные (в рамках принятой модели) соотношения [4] для вероятностей P при некоторых дополнительных допущениях могут быть упрощены и представлены в виде зависимостей от наглядного комплексного параметра – значения rV, определяющего математическое ожидание $\bar{l}$ количества случайных совпадений при проверке V вариаций по словарю $S(q,n,N)$. Примем, в частности допущение, что вероятность $P(l>1)$ появления более одного случайного (т.е. ложного) совпадения пренебрежимо мала. Это предположение близко к реальности для $r < 10^3 \div 10^{-4}$.

Действительно, в этом случае

$P(l=0) = (1-r)^V \approx 1 - rV$;

$P(l=1) = Vr(1-r)^{V-1} \approx rV$;

При этом $P(l=1)$ как раз совпадает со значением математического ожидания $\bar{l}$, т.к.

$\bar{l} = \sum_{l=1}^{V} l C_V^l r^l (1-r)^{V-l} = rV$.

В рамках принятого допущения значения P определяются следующими упрощенными выражениями (через P_Σ обозначена суммарная вероятность появления корректируемых ошибок, для которых генерируются вариации).

Алгоритм АКМВ

$P_{АК} \approx P_\Sigma(1-0{,}5rV)$; $P_{РК} \approx (1-P_\Sigma)(1-rV)$; $P_{ЛК} \approx rV(1-0{,}5P_\Sigma)$;

Алгоритм АОК

$P_{АК} \approx P_\Sigma(1-rV)$; $P_{РК} \approx P_\Sigma rV + (1-P_\Sigma)(1-rV)$; $P_{ЛК} \approx (1-P_\Sigma)rV$;

Алгоритм ПАКМВ

$P_{ПАК}(2) \approx P_\Sigma$; $P_{РК} \approx (1-P_\Sigma)$; $P_{ЛК} \approx 0$;

Алгоритм АПАКК

$P_{АК} \approx P_\Sigma(1-rV)$; $P_{ПАК}(2) \approx P_\Sigma rV$; $P_{РК} \approx (1-P_\Sigma)(1-rV)$; $P_{ЛК} \approx (1-P_\Sigma)rV$.

В табл. 3 приведены ориентировочные значения вероятностей P_i ошибок E_i [4] и соответствующие выражения для определения количества генерируемых вариаций V_i.

Табл.3

E_i	E_1	E_2	E_3	E_4	E_5
P_i	0,5557	0,1567	0,1204	0,0664	0,0322
V_i	$V_1 = (q-1)\cdot n$	$V_2 = n$	$V_3 = q\cdot(n+1)$	$V_4 = n-1$	$V_5 = (q-1)^2 \cdot C_n^2 - n + 1$

Степень отклонения приближенных выражений от точных иллюстрирует следующий пример. Для $E_{1,2,3,4}$, значения параметра $\alpha = 2$ (т.е. $n=8$, $r=10^{-4}$; $V=177$) и алгоритма АПАКК приближенные, значения оказываются равными: $P_{АК} \approx 8{,}832\cdot10^{-1}$, $P_{ПАК}(2) \approx 1{,}591\cdot10^{-2}$, $P_{РК} \approx 9{,}9\cdot10^{-2}$, $P_{ЛК} \approx 1{,}784\cdot10^{-3}$. Как видно, приближенные значения весьма близки к соответствующим данным табл. 1.

4. Практическое применение общего метода и, в частности, выбор алгоритма и ансамбля корректируемых ошибок зависит от многих факторов, в совокупности трудно поддающихся аналитическому учету. Среди них характеристика информации, т.е. словаря (значения N, n, q), вычислительные мощности ЭВМ (скорость генерации поиска и обработки вариаций, режим ввода и контроля – корректировки (on-line, off-line). Имитационное моделирование, проведенное на компьютере Celeron-1000/256MB с целью получения ориентировочных оценок скорости выполнения процесса, дало результаты, приведенные на графиках рис.1. Как видно из приведенных графиков и данных табл.3 существенных практических ограничений для автоматической и полуавтоматической корректировки "полного" ансамбля ошибок $E_{1,2,3,4,5}$ со стороны возможностей даже сравнительно слабого компьютера не имеется – даже при весьма больших объемах словаря и применении алгоритмов АОК и АПАКК, требующих полного перебора всех вариаций для подтверждения однозначности совпадения. Из этого следует, в частности, что для определенных условий (большие значения n, малые значения N и q) может оказаться возможным и целесообразным расширение ансамбля корректируемых ошибок за счет охвата "двойных" ошибок: $E_1 + E_2$, $E_1 + E_3$, $E_1 + E_4$. Отметим в связи с этим, что ошибка E_5 по существу представляет собой двойную ошибку $E_1 + E_1$.

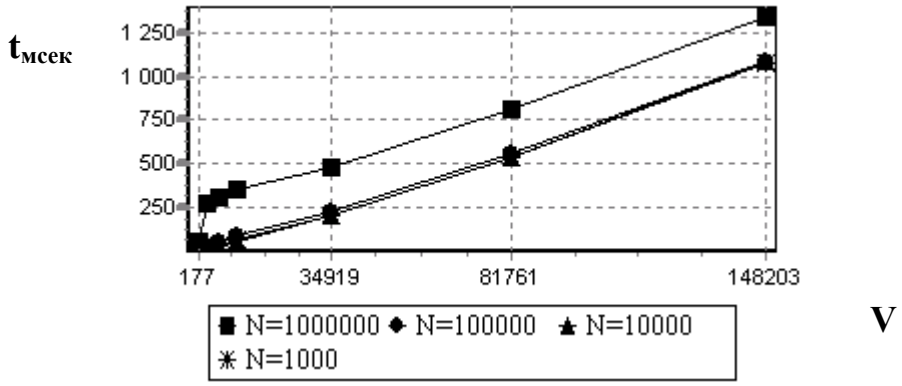

Рис.1

5. Приведенные данные и зависимости показывают, что общий метод и конкретные алгоритмы могут быть успешно использованы для снижения общей трудоемкости подготовки и ввода информации (в частности, формализованной) в ЭВМ. Выбор решений относительно конкретного алгоритма и ансамбля корректируемых ошибок для заданных параметров словаря (N,n,q) зависит от соответствующих значений r,V и технологического режима корректировки (on-line, off-line). Этот выбор можно очертить следующими ориентировочными рамками:

АКМВ - off-line, $r \le 10^{-7} \div 10^{-8}$, $t_{доп} < t(V) < 2t_{доп}$

АОК - off-line, $r = 10^{-3} \div 10^{-7}$, $t(V) < t_{доп}$

ПАКМВ – on-line, $r = 10^{-3} \div 10^{-7}$, $t_{доп} < t(V) < 2t_{доп}$

АПАКК - on-line, $r \le 10^{-7} \div 10^{-8}$, $t(V) < t_{доп}$.

Для $r > 10^{-3}$ применение общего метода нецелесообразно из-за относительно низких значений $P_{АК}, P_{ПАК}$ и, соответственно, высоких значений $P_{ЛК}, P_{РК}$.

Следует в заключение отметить, что в определенных случаях структура словаря может не соответствовать принятой модели. В этих случаях значения rV (или непосредственно P) для конкретного словаря могут быть определены путем предварительного "разового" моделирования и прямого перебора возможных частных исходов для конкретных ошибок.

ЛИТЕРАТУРА

1. AfterScan. http://www.afterscan.com/ru.

2. Бояринов И.М., Давыдов А.А., Мамедли Э.М., Смеркис Ю.Б. Использование помехоустойчивого кодирования для защиты информации от ошибок оператора. - М.: АТ, 1983.-№3.-С. 5-49.

3. Дремов И.В., Литвинов В.А. Автоматическая коррекция ошибок оператора на основе словаря-эталона // УсиМ.- 1989.- №3.-С.77-80.

4. Кузьменко Г.Є., Литвинов В.А., Майстренко С.Я., Ходак 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.