УДК 681.51:57

МЕТОД УЧЕТА ХАРАКТЕРИСТИК СЛОВАРЯ ПРИ АВТОМАТИЧЕСКОМ ОБНАРУЖЕНИИ И ИСПРАВЛЕНИИ ОШИБОК ПОЛЬЗОВАТЕЛЯ

В.А. Литвинов, С.Я. Майстренко
Інститут проблем математичних машин і систем НАН України
e-mail: takam@post.com

При анализе и ориентировочной оценке вероятностных характеристик методов обнаружения и исправления ошибок пользователя по словарю допустимых слов [1,2] принимается допущение о случайном распределении слов словаря в диапазоне всевозможных значений. При этом вероятность необнаружения ошибки определяется через параметр $r = N/q^n$, где N - количество реальных n-символьных слов словаря, а q - алфавит представления слов. Через r определяются и характеристики общей модели испытаний Бернулли в приложении к анализу характеристик автоматической идентификации и коррекции ошибок по словарю [2]. Так, вероятность $P(g,r,V)$ появления в точности g случайных совпадений V произвольных n-символьных комбинаций со словарем в моделях [2] определяется выражением:

$$P(g,r,V) = C_V^g r^g (1-r)^{V-g}.$$

Фактические распределения слов в реальных словарях могут отличаться от случайного и искажать результаты оценки в сторону заниженных оценок необнаруженных ошибок и значений $P(g,r,V)$. Сущность предлагаемого метода учета реальных характеристик конкретного словаря в моделях [2] заключается в использовании модели [1] для прямого определения некоторого эквивалентного значения $r_{экв}$, которое соответствует реальному относительному количеству необнаруживаемых ошибок и в соотношении с $r = N/q^n$ может служить мерой случайности распределения слов.

Примем следующие определения и обозначения:

S - полное множество n-символьных слов исследуемого словаря в алфавите q мощностью N;

T - тестовое подмножество S мощностью N_T, выбранное для исследования; для небольших словарей T может совпадать с S;

K_k - тестовый ансамбль специфических ошибок, включающий классы $E_1 \div E_L$, $k = 1,...,L$.

P_k - вероятность (относительная частота появления) ошибок класса E_k;

W_{kjT} - множество вариаций (искажений) слова T_j $(j = 1,...,N_T)$ в классе ошибок E_k;

V_{kjT} - мощность множества W_{kjT};

M_{kT} - подмножество вариаций, совпадающих со словарем T, мощностью N_{kT} ($M_{kT} = \bigcup_j M_{kjT}, N_{kT} = \sum_j N_{kjT}$).

Процесс анализа распределения слов в S состоит из следующих этапов:

- задание T (например, каждое 5-е, 10-е, 100-е и т.п. слово S);
- задание K_k (например, однократные транскрипции E_1 + пропуски символов E_2 + добавления символов E_3 + транспозиции E_4 и т.п.;
- генерации W_{kT};
- определение M_{kT} и подсчет N_{kT};
- подсчет относительного количества необнаруживаемых специфических ошибок для данного конкретного словаря, определяющего значение $r_{экв}$.

Выполнение последнего этапа основано на построении наборов всех вариаций (искажений типа E_k) каждого слова тестового подмножества T и проверка наличия соответствующей вариации в T. Нахождение вариации в T означает, что соответствующая конкретная ошибка не обнаруживается; суммарное (для всех слов T) количество найденных вариаций класса E_k составляет величину N_{kT}, а их относительное количество является частным от деления N_{kT} на произведение $V_{kT} \cdot N_T$.

Таким образом,

$$r_{экв} = P_1 \cdot \frac{N_{1T}}{V_{1T} \cdot N_T} + ... + P_L \frac{N_{LT}}{V_{LT} \cdot N_T} + P_{L+1} \frac{N}{q^n} = \frac{1}{N_T} \sum_{k=1}^{L} P_k \frac{N_{kT}}{V_{kT}} + P_{L+1} \frac{N}{q^n}, \qquad (1)$$

где $P_{L+1} = 1 - \sum_{k=1}^{L} P_k$.

Выбор T основан на ограничениях со стороны допустимого полного времени выполнения анализа словаря, т.е. чрезмерно большой словарь может быть "прорежен" и обработан выборочно. Критерием оценки для принятия решения здесь может служить суммарное количество C_T операций поиска в T, равное:

$$C_T = \sum_{k,j} V_{kjT} = N_T \sum_{k} V_{kT} .$$

Что же касается выбора W_{kT}, то здесь возможны два "крайних" случая [1].

1. Игнорирование специфических ошибок, т.е. предположение, что все ошибки носят случайный характер. Тогда $L = 0,\ r_{экв} = N/q^n$ и мы имеем "обычную" идеализированную оценку, соответствующую допущению о случайном распределении слов S.

2. Игнорирование случайных ошибок, как некоторого объединенного класса, т.е. детальный анализ всех классов ошибок, как специфических. В этом, практически нереальном случае, мы могли бы получить полную и максимально точную картину результативности процесса обнаружения всевозможных ошибок.

Поскольку редкие ошибки дают малый вклад в суммарное значение для $r_{экв}$ практически целесообразным представляется выбор такого значения L, для которого $\sum_{k=1}^{L} p_1 = 0.8 \div 0.9$. Это соответствует "полному" ансамблю специфических ошибок человека [2].

Пример расчета

j	Словарь S	N_{1jTi}			N_{1jT}	N^{p}_{2jTii}		N_{2jT}
		$i=3$	$i=2$	$i=1$		$ii=32$	$ii=21$	
1	0 0 1	1	0	4	5	0	0	0
2	0 0 2	3	1	4	8	0	0	0
3	0 0 3	0	1	4	4	0	0	0
4	0 0 5	1	0	4	5	0	0	0
5	0 0 7	0	1	4	4	0	0	0
6	0 1 2	1	1	3	5	1	0	1
7	0 1 3	0	1	3	4	0	0	0
8	0 1 7	0	1	3	4	0	0	0
9	0 1 9	0	0	3	3	0	0	0
10	1 0 2	3	1	3	7	1	0	1
11	1 0 4	0	1	3	4	0	0	0
12	1 0 5	1	0	3	4	0	0	0
13	1 0 8	0	1	3	4	0	0	0
14	1 1 0	0	0	3	3	0	0	0
15	1 1 2	1	1	3	5	0	0	0
16	1 1 4	0	1	3	4	0	0	0
17	1 1 8	0	1	3	4	0	0	0
18	2 0 2	3	1	0	4	0	0	0
19	7 0 1	1	0	1	2	0	0	0
20	7 0 2	3	1	1	5	0	0	0
	$\sum$	18	14	58	$N_{1T}=90$	2	0	$N_{2T}=2$

В таблице приведен пример расчета для фрагмента гипотетического словаря $S \equiv T$, $N = N_T = 20;\ \ q = 10;\ \ L = 2$; E_1 - однократные транскрипции, $P_1 = 0.555,\ \ V_{1jT} = 27;\ \ E_2$ - смежные транспозиции, $P_2 = 0.066,\ \ V_{2jT} = 2$. Расчет по выражению (1) дает значение $r_{экв} = 0.102$, т.е. примерно в 5 раз больше, чем идеализированное значение $r = N \big/ q^n = 2 \cdot 10^{-2}$.

Список литературы

1. Кузьменко Г.Е., Литвинов В.А., Литвинова А.Н., Майстренко С.Я. Модель анализа и оценки эффективности методов логического контроля информации // Математичні машини і системи.-2002.-№ 1.- С.49-55.
2. Литвинов В.А., Майстренко С.Я. Автоматическое исправление ошибок пользователя в человеко-машинных системах. Методы и характеристики // Матеріали науково-практичної конференції з міжнародною участю “Системи підтримки прийняття рішень. Теорія і практика”. - Київ: ІПММС НАНУ.-2005.- 7 червня. - С.112-115.