

УДК 681.3

**УПОРЯДОЧЕННИЙ ПЕРЕБОР ВАРИАЦІЙ ИСКАЖЕННОГО СЛОВА ПРИ
АВТОМАТИЧЕСКОЙ ИДЕНТИФИКАЦИИ ОШИБОК ПОЛЬЗОВАТЕЛЯ**

В.А. Литвинов, С.Я. Майстренко, И.Н. Оксанич

Институт проблем математических машин и систем НАН Украины

e-mail: maistr@immsp.kiev.ua

Процесс идентификации ошибочного слова и автоматического исправления ошибок пользователя при вводе данных [1], нечетком поиске образца и т.п. возможен на основе перебора вариаций искаженного слова и поиска в словаре-справочнике соответствующего ошибке допустимого значения [1].

Рассмотрим критерии и ожидаемую результативность упорядоченной стратегии генерации и перебора, направленной на уменьшение среднего количества перебираемых вариаций и, соответственно, на ускорение получения результата поиска.

Естественным интуитивным решением является генерация и перебор вариаций в порядке убывания вероятностей p_j появления ошибок пользователя. Неэффективность такого решения определяется существенной неравномерностью распределения количества вариаций V_j , которые требуется перебрать для ошибки класса E_j . Так, например, для идентификации однократной ошибки E_1 (вероятность $p_1 = 0.5557$ [1]) необходимо перебрать вариаций более, чем на порядок больше по сравнению с идентификацией вставки символа E_2 ($p_2=0.1567$) или транспозиции соседних символов E_4 ($p_4=0.0664$). Поэтому адекватная стратегия упорядоченного перебора должна базироваться на учете вероятностей ожидаемой результативности проверки одной вариации π_j , а не множества вариаций мощностью V_j . Такой подход попутно проясняет целесообразность рассмотрения и включения в ансамбль корректируемых ошибок и более редких ошибок, чем это принято в [1], - при условии, однако, что их идентификация требует перебора небольшого количества вариаций.

Прежде, чем получить количественные оценки результативности упорядоченного перебора покажем, что предлагаемый подход дает строго оптимальное решение для сокращения среднего числа переборов при идентификации корректируемой ошибки. Обозначим через $\bar{v}(y)$ среднее количество переборов вариаций в порядке y до нахождения в словаре совпадения вариации с допустимым словом при условии, что произошла именно и только ошибка из ансамбля корректирующих ошибок.

В соответствии с определением

$$\bar{v}(y) = \sum_{x=1}^V \pi_x(y) x, \quad (1)$$

где V - количество перебираемых вариаций;

$\pi_x(y)$ - удельная вероятность успешности одной вариации.

Оптимальное решение о порядке y перебора вариаций предлагает следующая простая, но важная теорема [3].

Пусть (a) и (b) есть две конечных системы (равновеликие множества) чисел.

Теорема 368. Если (a) и (b) заданы с точностью до перестановки, то $\sum ab$ наибольшее значение, когда (a) и (b) обе монотонно убывают или обе монотонно

возрастают, и наименьшее значение, когда одна из них монотонно возрастает, а другая – монотонно убывает.

Поскольку x в (1) монотонно возрастает, для минимизации $\bar{v}$ порядок y должен располагать $\pi_x(y)$ строго в порядке убывания (невозрастания).

Группируя слагаемые (1) с одинаковыми значениями $\pi_x(y)$, получим:

$$\bar{v}(y) = \sum_y \pi_y \left[V_y \sum_{k=1}^{y-1} V_k + \frac{V_y(V_y+1)}{2} \right], \quad (2)$$

$$\text{где } \pi_y \equiv \pi_x(y) = \frac{p_y}{V_y}.$$

В таблице приведены исходные и расчетные данные для расширенного (по сравнению с [1]) ансамбля корректируемых ошибок из 8 классов [4]: E_1 - однократные транскрипции, E_2 - добавление символа; E_3 - пропуск символа; E_4 - транспозиция смежных символов; E_5 - неспецифическая двукратная транскрипция; E_6 - двукратная транскрипция смежных идентичных символов; E_7 - транспозиция символов через одну позицию, E_8 - двукратная транскрипция идентичных символов, расположенных через одну позицию. В таблице приняты следующие дополнительные обозначения:

α - нормирующий множитель; q - алфавит символов слова;

n - длина слова (количество символов).

E_j	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8
p_j	0.5557	0.1567	0.1204	0.0664	0.0176	0.0081	0.0037	0.0028
$q=32; \quad n=8$								
V_j	248	8	288	7	26833	31	6	31
$\pi_j \cdot 10^4$	22.4	195.9	4.2	94.9	0.007	2.6	6.2	0.9
y_0	1	2	3	4	5	6	7	8
y_1	3	1	5	2	8	6	4	7
y_2	6	8	4	7	1	3	5	2
$\bar{v}(y_o)$	901.3							
$\bar{v}(y_1)$	411.3							
$\bar{v}(y_2)$	27025.2							
$q=10; \quad n=12$								
V_j	108	12	130	11	5271	9	10	9
$\pi_j \cdot 10^4$	51.5	130.6	9.3	60.4	0.033	9	3.7	3.1
y_0	1	2	3	4	5	6	7	8
y_1	3	1	4	2	8	5	6	7
y_2	6	8	4	7	1	3	5	2
$\bar{v}(y_o)$	236.1							
$\bar{v}(y_1)$	133.6							
$\bar{v}(y_2)$	5423.3							

Приведенные рассуждения и иллюстрирующие их расчетные данные дают основания для следующих выводов.

1. Упорядоченный по удельным вероятностям перебор вариаций искаженного слова заметно снижает объем перебора. Так, в идеализированном случае, когда произошла именно корректируемая ошибка (вероятность $\sum p_i = 0.9314$) и практическом отсутствии синонимов (избыточность словаря $r < 10^{-3} - 10^{-4}$ [1]), $\bar{v}(y_1)$ для оптимальной упорядоченности равно 411.3 и 133.6, $\bar{v}(y_0)$ для интуитивной стратегии перебора в порядке убывания p_j равно 901.3 и 236.1, а наихудший порядок перебора дает значения $\bar{v}(y_2)$, равное 27025.2 и 5423.3.
Это свойство существенно для алгоритмов 1 и 2 обработки словаря [1]. Естественно, примерно в 7% случаев, когда произошла некорректируемая ошибка, порядок перебора не играет никакой роли.
2. Несмотря на сравнительно малую вероятность появления ошибок класса $E_5 - E_8$, их целесообразно включать в ансамбль корректируемых ошибок, благодаря малой дополнительной «нагрузке» на общий объем перебора.

Литература

1. 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.
2. Ukrainian Context Optimizer. <http://www.uco.ua/infosection> 8 - doc 12.
3. Г.Е. Харди, Д.Е. Литтлвуд, Г.Полна. Неравенства. Государственное издательство Иностранной литературы. М., 1948г.
4. Литвинов В.А., Крамаренко В.В. Контроль достоверности и восстановления информации в человеко-машинных системах.-Киев: Техніка, 1986.-200 с.