УДК 681.3

Л.В. БЕЛОУС, В.А. ЛИТВИНОВ, С.Я. МАЙСТРЕНКО

МОДЕЛЬ УПРЕЖДАЮЩЕЙ ПОДСКАЗКИ В ИНТЕЛЛЕКТУАЛИЗОВАННОМ ИНТЕРФЕЙСЕ ПОЛЬЗОВАТЕЛЯ

Abstract: *One of methods "of small intellectualization" of the user interface – identification of a complete word on its part and anticipatory help to the user (or automatic restoration) is considered. It is constructed and investigated the model of the anticipatory help. The estimations of economy of the user's labour expeditures are received using this method during data input. The influence of truncation of the entered word on resulting reliability of the data is investigated, the appropriate quantitative estimations are resulted. It is marked, that the received estimations may be useful and in other supplements, for example, for automatic definition of language of the text under the analysis of initial symbols of the entered word.*
Key words: *user's mistakes, automatic correction, truth of data.*

Анотація: *Розглядається один з методів «малої інтелектуалізації» інтерфейсу користувача – ідентифікація повного слова по його частині і надання підказки користувачу (чи автоматичне відновлення). Побудована і досліджена модель підказки. Отримано оцінки економії трудозатрат користувача при використанні методу у процесі введення даних. Досліджено вплив усікання слова, що вводиться, на результатну вірогідність даних, наводяться відповідні кількісні оцінки. Відзначається, що отримані оцінки можуть виявитися корисними й в інших випадках, наприклад, для автоматичного визначення мови тексту за аналізом початкових символів слова, що вводиться.*
Ключові слова: *помилки користувача, автоматична корекція, вірогідність даних.*

Аннотация: *Рассматривается один из методов «малой интеллектуализации» интефейса пользователя – идентификация полного слова по его части и упреждающая подсказка пользователю (или автоматическое восстановление). Построена и исследована модель упреждающей подсказки. Получены оценки экономии трудозатрат пользователя при использовании метода в процессе ввода данных. Исследовано влияние усечения вводимого слова на результатную достоверность данных, приводятся соответствующие количественные оценки. Отмечается, что полученные оценки могут оказаться полезными и в других приложениях, например, для автоматического определения языка текста по анализу начальных символов вводимого слова.*
Ключевые слова: *ошибки пользователя, автоматическая коррекция, достоверность данных.*

1. Введение

Одним из общих направлений повышения уровня «малой интеллектуализации» интерфейса пользователя является совершенствование методов и технологических средств диалога пользователя с системой (сценария, меню, сообщений об ошибках, подсказок и т.п.). К подобным методам можно отнести, например, автоматическую идентификацию и коррекцию (АИК) ошибок пользователя [1, 2]. Наряду с методами [1, 2] существует и другая возможность помощи пользователю, заключающаяся в оперативном посимвольном анализе вводимых данных, их идентификации и подсказке (или автоматической подстановке) символов незавершенного слова в тех случаях, когда становится однозначно ясна последовательность недовведенных символов. Возможность таких ситуаций и реализацию подобной функции иллюстрируют, например, Интернет-браузеры при наборе пользователем ранее использовавшегося адреса, сохраненного в специальном справочнике.

Ставится задача анализа существенных свойств и оценки характеристик упомянутого метода, который в дальнейшем будем называть методом автоматической идентификации и восстановления (АИВ).

2. Постановка задачи

Содержательная постановка задачи заключается в построении и анализе модели процесса АИВ – модели, позволяющей сформулировать критерии и получить оценки результативности подсказки. Так же, как и для автоматической коррекции, информационной основой подсказки является естественная информационная избыточность, определяемая некоторым словарем – эталоном S, содержащим множество "разрешенных" слов, некоторое подмножество которого подлежит последующему вводу. Параметры словаря в целом и содержащихся в нем слов определяют искомые характеристики метода.

Для изложения формальной постановки задачи введем следующие исходные понятия и обозначения.

Без потери общности мы можем считать каждое из слов словаря A_j в алфавите q – целым числом в позиционной системе счисления с основанием q. Обозначим через a_{ij} значения i-го, начиная с младшего, символа слова $A_j\left(i=1,...,n;\ j=1,...,N\right)$.

Примем следующие допущения:

– значения слов A_j распределены случайно-равномерно среди всевозможных комбинаций символов $a_1...a_n$ в диапазоне $0 \div q^n - 1$;

– словарь S упорядочен по возрастанию значений A_j.

Назовем комбинацию $A_j(k)$ значений k старших символов $a_n...a_{n-i+1}...a_k$ **детерминантом** слова A_j, если в словаре отсутствуют дополнительные слова с совпадающей комбинацией значений данных символов. Это означает, что детерминант полностью идентифицирует конкретное полное слово.

Задача заключается в определении:

– зависимости между ожидаемой длиной детерминанта $\bar{k}$ и существенными характеристиками S (параметрами q, n, N);

– влияния сокращения вводимой длины слова с n до k на результатную достоверность ввода.

3. Алгоритм и модель АИВ

3.1. Алгоритм восстановления

Формальная схема алгоритма поиска $A_j(k)$ и восстановления A_j включает следующие этапы:

1. $i := n$.

2. Ввод (прием) i-го символа a_i^g очередного вводимого слова A^g.

3. Поиск в S ближайшего (в порядке возрастания значений) слова A_j, для которого $a_{n,j}a_{(n-1)j}...a_{(n-i+1)j} = a_n^g a_{n-1}{}^g ... a_{n-i+1}^g$. Если такого слова не существует, то переход к п. 5. Иначе переход к п. 4.

4. Если $a_{nj}a_{(n-1)j}...a_{(n-i+1)j} \neq a_{n(j+1)}a_{(n-1)(j+1)}...a_{(n-i+1)(j+1)}$, то $k := n-i+1$, комбинация символов $a_{nj}a_{(n-1)j}...a_{(n-i+1)j} := A_j^g(k)$ и остальные символы $a_{(n-i)j}...a_{1j}$ восстанавливаются автоматически.

Иначе $i := i+1$, переход к п. 2.

5. Вывод пользователю сообщения о наличии допущенной в A^g ошибки.

Пример фрагмента гипотетического словаря, содержащийся в табл. 1, иллюстрирует результаты роботы алгоритма для приведенных слов.

Таблица 1. Пример фрагмента словаря

A_j	$A_j(k)$	k
…	…	…
700125	70	2
712501	71	2
723627	72	2
745680	745	3
747560	747	3
761313	761	3
769053	769	3
773131	77	2
…	…	…

Содержательный смысл алгоритма заключается, как видно из его формального описания, в постепенном последовательном сужении области поиска детерминанта вводимого слова $(a_n \to a_n a_{n-1} \to a_n a_{n-1} a_{n-2} \to ...)$ до достижения искомого результата. После определения $A_j^g(k)$ восстановление слова A_j^g может быть зафиксировано автоматически или предложено пользователю для подтверждения, в зависимости от избранной технологии восстановления. П. 5 алгоритма означает, что искомый детерминант, а, значит, и полное слово, в словаре-эталоне отсутствует.

3.2. Модель анализа и оценки $\bar{k}$

Определим величину $l_j = A_{j+1} - A_j = \sum_{i=1}^{n} a_{i(j+1)} q^{i-1} - \sum_{i=1}^{n} a_{ij} q^{i-1}$ как текущий интервал между двумя произвольными значениями слов S. Очевидно, что $A_N = A_1 + \sum_{j=1}^{N-1} l_j$, а среднее значение $\bar{l} = \frac{A_N}{N-1}$. С учетом допущения о равномерном распределении *N* значений словаря среди

q^n значений возможных комбинаций $a_1 ... a_n$, пренебрегая разностями $[(q^n - 1) - A_N]$ и $(A_1 - 0)$, положим $\bar{l} \approx \frac{q^n}{N} = \frac{1}{r}$.

Легко показать, что в случае регулярной структуры словаря S с постоянным интервалом l_0 значения $\bar{\bar{k}} = k_0$ были бы одинаковыми для всех слов и определялись бы неравенством $q^{n-k_0} \le l_0 < q^{n-k_0+1}$.

При этом для $l_0 = q^m (m = 0,1,...,n-1)$ мы имеем $\bar{\bar{k}} = k_0 = n - m$.

Для получения приближенных оценок $\bar{k}$ при равномерном распределении рассмотрим непрерывную аппроксимацию словаря и процесса поиска детерминанта. Сопоставим диапазону всевозможных значений слов словаря отрезок прямой $0 \div q^n$ с текущей координатой x, а каждому реально существующему слову A_j – точку з координатой X_j. Будем интерпретировать распределение значений текущих интервалов как распределение интервалов ожидания l в случайном пуассоновском потоке событий с интенсивностью $1/\bar{l} = \frac{N}{q^n} = r$. Иными словами, дискретным значением l_j поставим в соответствие непрерывные значения l с распределением $P(l) = cr\,exp(-rl)$, где c – нормирующий множитель. Поставленную задачу поиска $\bar{k}$ в терминах принятой аппроксимации можно трактовать следующим образом.

По дороге мимо наблюдателя проезжают автомобили, распределение которых во времени подчиняется закону Пуассона, следовательно, интервалы времени между последовательными проездами двух автомобилей распределены экспоненциально. Водителю автомобиля, время ожидания которого находится в диапазоне $(l_{t-1} = q^{t-1}) \div (l_t = q^t)$, наблюдателем выдается $k = n - t + 1$ единиц "вознаграждения" $(t = 1...n)$. Водителям, время ожидания которых превышает $l_n = q^n$, не выдается ничего. Необходимо найти среднее значение вознаграждения.

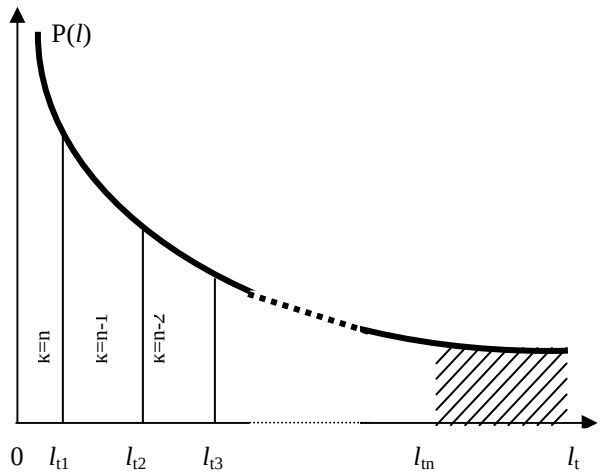

Рис. 1. Распределение $P(l)$

Из рисунка, иллюстрирующего эти положения, видно, что относительное количество водителей, получающих вознаграждение в $(n-t+1)$ единиц, соответствует вероятности

$$P\{l_{t-1} \le l \le l_t\} = P_t = \int_{l_{t-1}}^{l_t} P(l)dl = cr \int_{l_{t-1}}^{l_t} \exp(-rl)dl .$$

Нормирующий множитель c определяет вероятность события $l < q^n$ (т.е. относительную часть водителей, которые получают хоть какие-то вознаграждения). Следовательно, после интегрирования получим

$$P_t = \frac{\exp(-rq^t) - \exp(-rq^{t-1})}{1-\exp(-rq^n)} \qquad \text{и} \qquad (1)$$

$$\bar{k} = \sum_{t=1}^{n} (n-t+1) \cdot P_t . \qquad (2)$$

После некоторых промежуточных преобразований (1), (2), в частности, перегруппировок слагаемых в порядке увеличения “вознаграждения“, получим:

$$\bar{k} = \frac{\sum_{t=1}^{n} t\left[\exp(-rq^{n-t}) - \exp(-rq^{n-t+1})\right]}{1-\exp(-rq^n)} . \qquad (3)$$

В выражении (3) величина в квадратных скобках числителя определяет вероятность того, что некоторый текущий интервал попадает в диапазон ($q^{n-t} \div q^{n-t+1}$), в этом случае $k = t$.

Нормирующий знаменатель учитывает тот факт, что $l < q^n$ и, следовательно, “хвост“ теоретической кривой $P(l) = r\,exp(-rl)$, лежащий правее точки $l = q^n$, определяет вероятность нереальных событий.

В табл. 2 приведены значения $\bar{k}$, рассчитанные по выражению (3) для $q = 10$, $n = 8{,}6$.

Таблица 2. Значения $\bar{k}$ для $q = 10$, $n = 8{,}6$.

r / n	10^{-3}	10^{-4}	$5 \cdot 10^{-5}$	10^{-5}	$5 \cdot 10^{-6}$	10^{-6}	10^{-7}
8	5,7292	4,7374	4,4406	3,7382	3,4410	2,7383·10	1,7384
6	3,7312	2,7376	2,4407	1,7383	1,4508	1,1680·10	1,1162

4. Оценка достоверности ввода информации

Устанавливая зависимость достоверности ввода информации от параметра k, примем во внимание следующее. С одной стороны, при вводе неполного слова уменьшается количество ошибок, допущенных пользователем, за счет меньшего количества вводимых символов. Это уменьшение пропорционально отношению n/k. С другой стороны, при контроле сравнением неполных слов уменьшается количество обнаруживаемых ошибок (п. 5 алгоритма) за счет

уменьшения интервала между разрешенными усеченными словами и соответствующего уменьшения относительного количества разрешенных значений неполных слов. Для оценки совместного действия противоположно направленных тенденций введем в рассмотрение параметр

$$\eta = \frac{n[1-d(n)]}{k[1-d(k)]},$$

где $d(n)$ и $d(k)$ – соответственно достоверность информации при контроле по полному *n*-символьному слову и усеченному *k*-символьному словам.

Значение n определяет относительное количество ошибок на выходе контроля по полному n-символьному слову по сравнению с усеченным k-символьным.

При принятом допущении о равномерном распределении N значений слов словаря в интервале $0 \div q^n$ величина $d(n) = 1 - N/q^n = 1 - r$. Для получения зависимости $d(k)$ воспользуемся соотношениями известной задачи о размещениях [3]. Как известно, вероятность $P[x, y, c(o)]$ того, что при случайных бросаниях X шаров («дробинок» [4]) в Y ящиков C(0) ящиков останутся пустыми, равна

$$P[x, y, c(0)] = \exp(-\phi) \cdot \phi^{c(0)} / c(0)!\ ,$$

где

$$\phi = y \exp(-x/y)\ .$$

Для оценки приближенного значения математического ожидания $\bar{c}(0)$ воспользуемся результатами [4]:

$$\bar{c}(0) \approx exp(-x/y) \cdot (y - x/2y)\ .$$

Поставим в соответствие x шарам N слов словаря, а y ящикам – q^k потенциальных всевозможных значений усеченных k-символьных слов. В этом случае, с учетом прежнего предположения о равномерном характере распределения значений k символов N слов словаря, для относительного количества обнаруженных ошибок $d(k)$ можно записать следующее достаточно очевидное выражение:

$$\mathrm{d(k)} = \frac{\bar{\mathrm{c}}(0)}{\mathrm{q^k}}.$$

Действительно, обнаружению ошибки в усеченном слове соответствует ситуация, когда после бросания x шаров наугад выбранный ящик оказывается пустым. При этом величина $\bar{C}(0)/q^k$ представляет собой относительное количество таких ситуаций.

Окончательно

$$\eta = \frac{nr}{k\left\{1 - \frac{1}{q^k} \cdot \left[\exp(-r(k)) \cdot (q^k - \frac{r(k)}{2})\right]\right\}}\ , \qquad (4)$$

где $r(k) = \frac{N}{q^k}$.

В табл. 3 в качестве примера приведены значения η для $q = 10$; $N = 1000$; $r = 10^{-3}; 10^{-5}$.

Таблица 3. Значение η для выбранных значений параметров q, N, r

k	2	3	4	5	6	7	8
η ($r=10^{-3}$, $n=6$)	$3.000 \cdot 10^{-3}$	$3.163 \cdot 10^{-3}$	$1.576 \cdot 10^{-2}$	$1.206 \cdot 10^{-1}$	1.000	–	–
η ($r=10^{-5}$, $n=8$)	$4.000 \cdot 10^{-5}$	$4.217 \cdot 10^{-5}$	$2.101 \cdot 10^{-4}$	$1.608 \cdot 10^{-3}$	$1.334 \cdot 10^{-2}$	$1.142 \cdot 10^{-1}$	1.000

5. Выводы

Анализ выражений (3), (4) и данных табл. 2, 3 позволяет сделать следующие выводы:

1. Упреждающая подсказка в режиме АИВ позволяет получить существенное сокращение трудозатрат при вводе данных. Например, при вводе последовательности слов с характеристиками словаря $n = 8$, $N = 1000$ в среднем достаточно ввести меньше 4-х символов для идентификации и автоматического восстановления каждого 8-символьного слова, т.е. трудозатраты в этом случае сокращаются более, чем в 2 раза. Для конкретного случая относительное снижение трудоемкости, равное $\bar{k} / n$, определяется выражением (3).

2. Величина $\bar{k}$ по мере уменьшения r уменьшается при фиксированном n в логарифмическом масштабе практически линейно, стремясь к величине $log_q N$ при $r \to 0$. Это свойство интуитивно очевидно: чем ближе r к нулю (т.е. чем больше избыточность и чем больше текущий интервал l между реальными словами), тем меньше требуется символов для идентификации всего слова – вплоть до минимально необходимого количества символов для представления N слов в алфавите q. При этом, как можно видеть из табл. 2, значения $\bar{k}$ близки к значениям k_o, присущим регулярной структуре. Поэтому простое соотношение для k_o может быть использовано для ориентировочной оценки $\bar{k}$ при аппроксимации реальной равномерной структуры словаря некоторой «близкой» регулярной структурной.

3. И восстановление полного слова, вводимого в компьютер по детерминанту, и контроль информации по словарю-эталону используют одну и ту же избыточность и один и тот же базовый инструмент (т.е. словарь) Поэтому ввод в режиме АИВ в значительной мере снижает эффективность контроля по словарю. Как следует из анализа выражения (4), иллюстрированного

данными табл. 3, чем выше результативность подсказки (чем меньше r и, следовательно, $\bar{k}$), тем хуже характеристики достоверности выходной информации (больше ошибок на выходе). Поэтому режим АИВ целесообразно применять в сочетании с иными способами контроля.

В заключение отметим, что в перспективе изложенный подход к оценке $\bar{k}$ может оказаться полезным и в некоторых других приложениях, – например, при автоматическом определении языка текста по анализу начальных символов вводимого слова, автоматическом восстановлении искаженного слова по словарю и т.п.

СПИСОК ЛИТЕРАТУРЫ

1. Дремов И.В., Литвинов В.А. Автоматическая коррекция ошибок оператора на основе словаря-эталона // УсиМ. – 1989. – № 3. – С. 77 – 80.
2. Литвинов В.А., Майстренко С.Я., Ступак Н.Б. Некоторые оценки вероятностных характеристик процесса автоматической идентификации ошибок пользователя на основе эталонного словаря // УсиМ. – 2001. – № 2. – С. 21 – 24.
3. Феллер В. Введение в теорию вероятности и ее приложения. – М.: Мир, 1967. – Т. 1. – С. 109 – 110.
4. Севастьянов Б.А., Чистяков В.П. Асимптотическая нормальность в классической задаче о дробинках // Теория вероятностей и ее применения. – 1964. – Т. 9. – Вып. 2. – С. 233 – 237.