УДК 658.52:655

С.И. Карпович, доцент; Д.С. Карпович, ассистент

ОПРЕДЕЛЕНИЕ ХАРАКТЕРИСТИК КОРРЕКТОРА ДЛЯ ОБЕСПЕЧЕНИЯ ЗАДАННОГО КАЧЕСТВА КОНЕЧНОГО ПРОДУКТА

In paper the idea of a determination of a matrix of passage for operations of a proofreading of the text is considered, radiating from given vectors of states of the text before operation of a proofreading. For this purpose the method of discrete Markovian processes is used. In outcome the algorithm of a determination of a matrix of passage for any size n of vectors of a state of the text is offered.

Одной из важнейших характеристик продукции является ее качество. В случае рассмотрения текста книги, статьи, да и любого издания основной характеристикой текста является количество ошибок, которые рассматриваемый текст содержит. Для анализа качества текста целесообразно применять метод дискретных марковских процессов. В этом случае вектор состояния текста определяется в виде матрицы-строки:

$$\overline{P} = \begin{bmatrix} P_0 & P_1 & P_2 & \dots & P_n \end{bmatrix}, \qquad (1)$$

где P_i, — вероятность того, что количество ошибок в тексте равно i.

Элементы вектора состояния удовлетворяют следующему очевидному условию:

$$\sum_{i=0}^{n} P_i = 1. \qquad (2)$$

Преобразование текста после каждого этапа технологического процесса описывается матрицей перехода:

$$P = \begin{bmatrix} P_{00} & P_{01} & P_{02} & \cdots & P_{0n} \\ P_{10} & P_{11} & P_{12} & \cdots & P_{1n} \\ P_{20} & P_{21} & P_{22} & \cdots & P_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ P_{n0} & P_{n1} & P_{n2} & \cdots & P_{nn} \end{bmatrix}. \qquad (3)$$

Элементы матрицы P_{ij} означают вероятность перехода текста из состояния i в состояние j, т. е. вероятность того, что число ошибок изменится от i до j. Состояние n будем считать таким, когда число ошибок в тексте $i \geq n$.

Сумма элементов матрицы в каждой строке равна единице:

$$\sum_{j=0}^{n} P_{ij} = 1. \qquad (4)$$

Для нахождения вектора состояния текста после его преобразования необходимо воспользоваться матричным произведением

$$\overline{P(1)} = \overline{P(0)} \cdot P, \qquad (5)$$

где $P(0)$ и $P(1)$ —строки, соответствующие векторам состояния текста до и после его преобразования.

При вводе текстовой информации в ЭВМ с помощью клавиатуры матрица кодирования приобретает треугольную форму:

$$P_{\kappa o \partial}=\begin{bmatrix} P_{00} & P_{01} & P_{02} & \cdots & P_{0n} \\ 0 & P_{11} & P_{12} & \cdots & P_{1n} \\ 0 & 0 & P_{22} & \cdots & P_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \quad (6)$$

Тогда уравнение (5) может быть записано в развернутом виде:

$$[PP_0 \quad PP_1 \quad PP_2 \quad \dots \quad PP_n]=[PD_0 \quad PD_1 \quad PD_2 \quad \dots \quad PD_n]\cdot\begin{bmatrix} P_0 & P_1 & P_2 & \cdots & P_n \\ 0 & P_0 & P_1 & \cdots & P_{n-1}+P_n \\ 0 & 0 & P_0 & \cdots & P_{n-2}+P_{n-1}+P_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \quad (7)$$

Данное матричное уравнение может быть записано в виде системы линейных уравнений:

$$\begin{cases} PD_0\cdot P_0=PP_0\ ; \\ PD_0\cdot P_1+PD_1\cdot P_0=PP_1\ ; \\ PD_0\cdot P_2+PD_1\cdot P_1+PD_2\cdot P_0=PP_2\ ; \\ \vdots \\ PD_0\cdot P_n+PD_1\cdot P_{n-1}+\ldots+PD_{n-1}\cdot P_1+PD_n\cdot P_0=PP_n\ . \end{cases} \quad (8)$$

Для проверки можно воспользоваться уравнением $\sum_{j=0}^{n} P_j=1$.

Для устранения ошибок в набранном тексте используются операции корректуры. Матрица корректуры в общем случае имеет вид, описываемый выражением (6). Если же корректор абсолютно грамотен, т. е. не вносит дополнительных ошибок в текст, то матрица корректуры имеет вид

$$P_{\kappa o p}=\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ P_{10} & P_{11} & 0 & \cdots & 0 \\ P_{20} & P_{21} & P_{22} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ P_{n0} & P_{n1} & P_{n2} & \cdots & P_{nn} \end{bmatrix}. \quad (9)$$

Рассмотрим идею определения коэффициентов в матрице корректуры для получения текста с заданным количеством ошибок PP при условии, что характеристики исходного текста PD также известны. Решение такой задачи «в лоб» не представляется возможным, так как в этом случае мы будем иметь дело с системой из $2n$ уравнений и оперировать при этом $n(n—1)/2+n$ неизвестными. Для нахождения искомых коэффициентов предлагается следующий алгоритм нахождения коэффициентов матрицы перехода для корректуры.

1. Рассмотрим матрицу размера 3x3. Тогда уравнение (5) примет вид

$$[PP_0 \quad PP_1 \quad PP_2]=[PD_0 \quad PD_1 \quad PD_2]\cdot\begin{bmatrix} P_{00} & 0 & 0 \\ P_{10} & P_{11} & 0 \\ P_{20} & P_{21} & P_{22} \end{bmatrix}.$$

$$\begin{cases} PD_0 \cdot P_{00} + PD_1 \cdot P_{10} + PD_2 \cdot P_{20} = PP_0 \ ; \\ PD_0 \cdot 0 + PD_1 \cdot P_{11} + PD_2 \cdot P_{21} = PP_1 \ ; \\ PD_0 \cdot 0 + PD_1 \cdot 0 + PD_2 \cdot P_{22} = PP_2 \ ; \end{cases} \tag{10}$$

$$\begin{cases} P_{00} = 1 \ ; \\ P_{10} + P_{11} = 1 \ ; \\ P_{20} + P_{21} + P_{22} = 1 \ . \end{cases} \tag{11}$$

Таким образом, имеем систему из шести уравнений с шестью неизвестными, которые можно найти численными или аналитическими методами. Найденные коэффициенты матрицы P в дальнейшем обозначим с индексом $н$.

2. Рассмотрим матрицу размера 4x4. Тогда уравнение (5) примет вид

$$\begin{bmatrix} PP_0 & PP_1 & PP_2 & PP_3 \end{bmatrix} = \begin{bmatrix} PD_0 & PD_1 & PD_2 & PD_3 \end{bmatrix} \cdot \begin{bmatrix} P_{00н} & 0 & 0 & 0 \\ P_{10н} & P_{11н} & 0 & 0 \\ P_{20} & P_{21} & P_{22} & 0 \\ P_{30} & P_{31} & P_{32} & P_{33} \end{bmatrix}.$$

$$\begin{cases} PD_0 \cdot P_{00н} + PD_1 \cdot P_{10н} + PD_2 \cdot P_{20} + PD_3 \cdot P_{30} = PP_0 \ ; \\ PD_0 \cdot 0 + PD_1 \cdot P_{11н} + PD_2 \cdot P_{21} + PD_3 \cdot P_{31} = PP_1 \ ; \\ PD_0 \cdot 0 + PD_1 \cdot 0 + PD_2 \cdot P_{22} + PD_3 \cdot P_{32} = PP_2 \ ; \\ PD_3 \cdot P_{33} = PP_3 \ ; \end{cases} \tag{12}$$

$$\begin{cases} P_{20} + P_{21} + P_{22} = 1 \ ; \\ P_{30} + P_{31} + P_{32} + P_{33} = 1 \ . \end{cases} \tag{13}$$

Таким образом, имеем систему 6 уравнений с 7 неизвестными. Для нахождения воспользуемся уравнением

$$\frac{P_{20i} - P_{21i}}{P_{21i} - P_{22i}} = \frac{P_{20} - P_{21}}{P_{21} - P_{22}}.$$

Данное соотношение имеет следующий физический смысл: в новой матрице P размера 4x4 числа в третьей строке должны находиться в той же пропорции, что и числа в третьей строке матрицы P размера 3x3.

Аналогичным образом рассматриваются матрицы с размером на одну строку и один столбец большие по сравнению с рассмотренными ранее матрицами, дополняя при этом систему уравнений соотношениями, аналогичными (14) до тех пор, пока размер не достигнет заданной величины n.

ЛИТЕРАТУРА

1. Ефимов М.В. Автоматизированное управление полиграфическим производством. – М.: Мир книги, 1998.
2. Технические средства переработки текста и иллюстраций / Под ред. М.В. Ефимова. – М.: Мир книги, 1994.
3. Берлин А.С., Воскресенский М.И., Ефимов М.В. и др. Автоматизированные системы переработки текстовой информации. – М.: Книга, 1991.