

УДК 519.772

С.В. Доценко, проф., д-р физ.-мат. наук, О.С. Зиборова

Севастопольский национальный технический университет

Студгородок, г. Севастополь, Украина, 99053

E-mail: root@sevgtu.sebastopol.ua

АНАЛИЗ ИНФОРМАЦИОННЫХ ХАРАКТЕРИСТИК ДИСКРЕТНЫХ СЛУЧАЙНЫХ ПРОЦЕССОВ

Создано программное средство для расчета информационных характеристик дискретных случайных процессов. Программа апробирована на текстовых файлах. Получены зависимости оценок энтропии и взаимной информации от длины текста, а также от числа символов, на которое один текст сдвигается относительно другого.

В настоящее время исследование случайных процессов в информационных системах в основном производится спектральным и корреляционным методами. Поскольку текст можно рассматривать как случайный процесс [1], в котором слова и буквы являются случайными дискретными величинами, то была поставлена задача разработки программного средства, позволяющего выполнять информационный анализ текстов. Преимущество информационного анализа по сравнению с корреляционным и спектральным состоит в том, что результаты первого не зависят от вида кодирования отдельных символов, чего нельзя сказать о последних видах анализа.

Для решения поставленной задачи была разработана рабочая программа на языке Object Pascal для расчета энтропии и взаимной информации двух текстов. Она позволяет проводить в интерактивном режиме информационные исследования одного или двух текстов на основе экспериментально-расчетных данных о параметрах, влияющих на количество информации. С помощью этой программы можно анализировать произвольные тексты на естественных и искусственных языках, в том числе и на языках программирования. Это могут быть два различных текста либо один и тот же текст, но сдвинутый относительно начальной позиции на некоторое число символов. Программа считывает тексты из файлов с расширением *.txt. При этом их ASCII-коды сохраняются в переменных-массивах.

В случае исследований одного и того же текста также создаются два массива: один – для исходного текста, а другой – для смещенного текста. Сдвиг смещенного текста осуществляется программно влево на заданное пользователем количество символов. Сдвиг текста вправо дает те же результаты, так как в данном случае от изменения направления сдвига меняются местами лишь массивы исходного и смещенного текстов.

Взаимная информация двух текстов X и Y определяется матрицей двумерных вероятностей

$$P_{xy} = \begin{bmatrix} P(x_1, y_1) & P(x_1, y_2) & \dots & P(x_1, y_m) \\ P(x_2, y_1) & P(x_2, y_2) & \dots & P(x_2, y_m) \\ \dots & \dots & \dots & \dots \\ P(x_n, y_1) & P(x_n, y_2) & \dots & P(x_n, y_m) \end{bmatrix}, \quad (1)$$

где $P(x_i, y_j)$ – двумерная вероятность ситуации, при которой в некоторой позиции текста, определенного на множестве символов X , встречается символ x_i , а в соответствующей позиции текста, определенного на множестве символов Y , – символ y_j .

Для определения двумерной вероятности $P(x_i, y_j)$ необходимо вести подсчет числа комбинаций $k(x_i, y_j)$ и подсчет общего числа символов N в тексте.

Тогда оценка вероятности появления пары символов рассчитывается по следующей формуле:

$$P(x_i, y_j) = \lim_{N \rightarrow \infty} \frac{k(x_i, y_j)}{N}.$$

Если длины исследуемых текстов отличаются, то число символов в обоих текстах принимается равным минимальному из них. Это касается исследований, проводимых как на одном, так и на двух разных текстах.

Таким образом, матрица (1) отражает вероятности пар символов, которые стоят в текстах на одинаковых позициях.

Наиболее интересными представляются исследования, выполненные с одним и тем же текстом. Они отражают связи между смежными и отдаленными символами. Об этом можно судить по матрице (1), изменяя число символов, на которое сдвигается текст относительно исходной позиции. При изменении сдвига изменяется объем взаимной информации. Можно также анализировать изменение объема взаимной информации в зависимости от величины выборки, взятой из данного текста.

От матрицы двумерных вероятностей (1) нетрудно перейти к одномерным вероятностям при помощи следующих формул

$$\left. \begin{array}{l} \sum_{i=1}^m P(x_1, y_i) = P_x(x_1) \\ \dots \\ \sum_{i=1}^m P(x_n, y_i) = P_x(x_n) \end{array} \right\}; \quad \left. \begin{array}{l} \sum_{i=1}^n P(x_i, y_1) = P_y(y_1) \\ \dots \\ \sum_{i=1}^n P(x_i, y_m) = P_y(y_m) \end{array} \right\},$$

где $P_x(x_i)$ – априорная вероятность появления символа x_i в первом текстовом сообщении, т.е. вероятность того, что случайный символ примет конкретное значение x_i ; $P_y(y_j)$ – вероятность появления символа y_j во втором тексто-

вом сообщении, т.е. вероятность того, что в нем случайный символ примет конкретное значение y_j .

Зная вероятности $P_X(x_i)$ и $P_Y(y_j)$, можно определить энтропию $H(X)$ множества символов X , энтропию $H(Y)$ множества символов Y , энтропию $H(X,Y)$ множества совместных символов X, Y , пользуясь соответственно следующими формулами [1-4]:

$$H(X) = - \sum_{i=1}^n P_X(x_i) \log_2 P_X(x_i),$$

$$H(Y) = - \sum_{j=1}^m P_Y(y_j) \log_2 P_Y(y_j),$$

$$H(X,Y) = - \sum_{i=1}^n \sum_{j=1}^m P(x_i, y_j) \log_2 P(x_i, y_j).$$

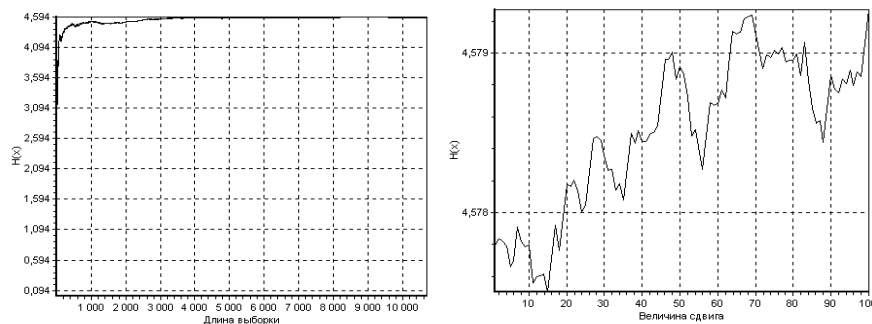
Средняя взаимная информация I_{XY} связана с $H(X)$, $H(Y)$ и $H(X,Y)$ следующим соотношением:

$$I_{XY} = H(X) + H(Y) - H(X,Y).$$

С помощью найденных соотношений было исследовано влияние длины выборки текста и величины сдвига между выборками текста на значение взаимной информации.

На основании экспериментальных данных были получены зависимости величин $H(X)$, $H(Y)$, $H(X,Y)$, $I(X,Y)$ от длины выборки русскоязычного текста и от величины сдвига выборки текста. Графики указанных зависимостей показаны на рисунках 1-4, где длина и сдвиг выборки определяется числом символов, а энтропии и информации измеряются в битах.

Из рисунка 1,а следует, что при увеличении длины выборки текста её энтропия стремится к некоторому постоянному значению, которое не превышает 4,6 бит. Полученное значение энтропии подтверждает известный



а) б)
Рисунок 1 – Зависимости энтропии $H(X)$: а) от длины выборки текста;
б) от величины сдвига выборки текста

факт, что для представления текста на русском языке в цифровом виде достаточно пятиразрядного двоичного кода. Стремление $H(X)$ к определенному пределу при увеличении длины выборки свидетельствует о наличии в текстах стохастической устойчивости. Если при каждом сдвиге выборки текста длиной 10000 символов рассчитывать значение его энтропии, то получим график, изображенный на рисунке 2, б. Если учесть различие масштабов по оси ординат двух графиков (рисунок 2), то видно, что с увеличением сдвига энтропия несколько возрастает с малозначительными колебаниями.

Зависимости совместной энтропии $H(X,Y)$ между исходной выборкой текста и той же выборкой, смещенной на один символ, от длины выборки показана на рисунке 2, а. Зависимость совместной энтропии $H(X,Y)$ между исходной выборкой длиной 10000 символов и той же выборкой, сдвинутой на некоторое число символов, от величины сдвига, изображена на рисунке 2, б. Как и в случае собственной энтропии, при увеличении длины выборки совместная энтропия стремится к некоторому постоянному значению, которое превышает собственную энтропию приблизительно в полтора раза (рисунок 2, а).

При увеличении сдвига тексты исходной выборки X и сдвинутой Y все меньше статистически связаны друг с другом. Поэтому при увеличении сдвига совместная энтропия стремится к сумме энтропий выборок $H(X,Y) = H(X) + H(Y) = 2H(X)$, что и подтверждается сравнением графиков, показанных на рисунках 1, а и 2, б.

Была найдена зависимость взаимной информации $I(X,Y)$ между исходной выборкой текста и той же выборкой, смещенной на один символ, от длины выборки текста (рисунок 3, а). Из полученных данных следует, что количество взаимной информации убывает и стремится к единице. Таким образом, при увеличении длины выборки величина $I(X,Y)$ стремится к сво-

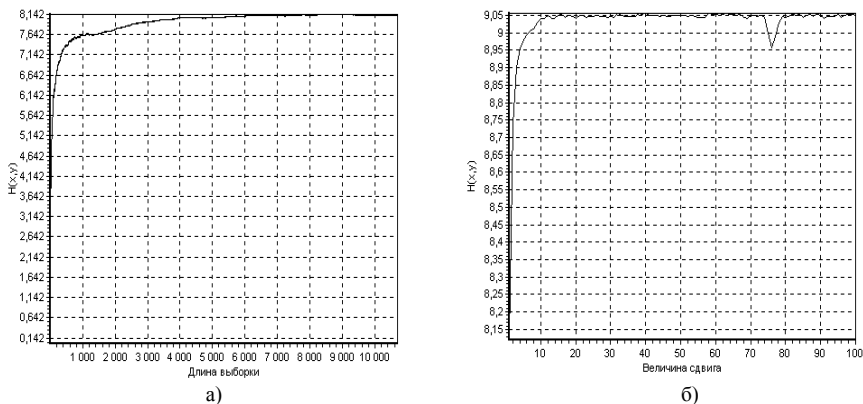


Рисунок 2 – Зависимость взаимной энтропии $H(X,Y)$: а) от длины выборки; б) от величины сдвига текста

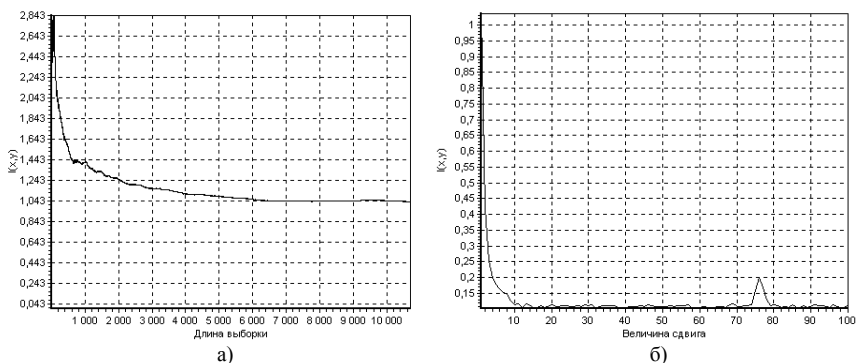


Рисунок 3 – Зависимость взаимной информации $I(X,Y)$: а) от длины выборки; б) от величины сдвига текста

ему математическому ожиданию. Так как выборки текста сдвинуты друг относительно друга на один символ, то при увеличении числа символов в выборке $N \rightarrow \infty$ величина $I(X,Y)$ стремится к единице.

Была получена зависимость взаимной информации между исходной выборкой длиной 10000 символов и той же выборкой, сдвинутой на некоторое число символов, от величины сдвига (рисунок 3,б) видно, что при увеличении сдвига взаимная информация быстро убывает, стремясь к нулю. Иными словами, в русском языке взаимная информация между символами имеет существенную величину при сдвиге не более чем на 5-6 символов.

При сдвиге текста на 76 символов (длину одной строки текста) зависимость совместной энтропии $H(X,Y)$ от величины сдвига текста имеет выброс (рисунки 2,б и 3,б), что объясняется повторением символа перевода каретки в конце каждой строки текста.

Все полученные выше графические зависимости и связанные с ними выводы относились к исследованиям одного и того же текста.

С помощью разработанной программы был выполнен также анализ двух различных текстов и получена зависимость взаимной информации этих текстов от длины выборки, которая изображена на рисунке 4. Видно, что при увеличении длины выборки текста взаим-

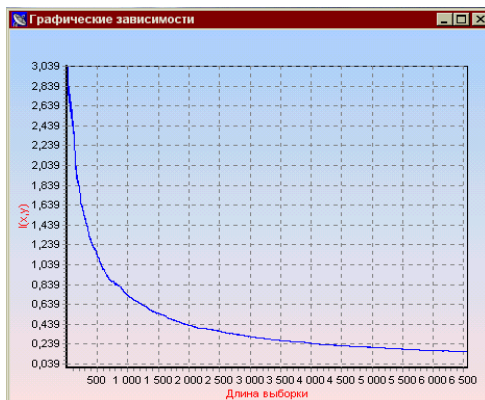


Рисунок 4 – Зависимость взаимной информации $I(X,Y)$ от длины выборки при исследовании двух текстов

ная информация монотонно убывает и для бесконечно длинного тексте стремится к нулю.

Полученные результаты позволяют оценить информационные характеристики текстовых документов на любых языках и могут быть полезными при определении предельных возможностей сжатия дискретной информации в процессе ее кодирования в цифровом виде, передачи, приема и архивации с целью хранения в памяти ЭВМ или на носителях информации.

При дальнейшем исследовании необходимо проанализировать текстовые документы на английском языке, а также тексты программ, написанных на различных языках программирования.

Библиографический список

1. Вентцель Е.С. Теория вероятностей / Е.С. Вентцель. — М.: Высш. шк., 2001. — 575 с.
2. Дмитриев В.И. Прикладная теория вероятности / В.И. Дмитриев. — М.: Высш. шк., 1989. — 320 с.
3. Колесник В.Д. Курс теории информации / В.Д. Колесник, Г.Ш. Полтырев. — М.: Наука, 1982. — 416 с.
4. Шеннон К. Работы по теории информации и кибернетике / К. Шеннон. — М.: Иностран. лит., 1963. — 829 с.

Поступила в редакцию 29.03.2004 г.