ПЕДАГОГИЧЕСКАЯ ИНФОРМАТИКА

СИМВОЛЬНОЕ ПРЕДСТАВЛЕНИЕ ИНФОРМАЦИИ И ЕЕ ИЗМЕРЕНИЕ

В.А. Бубнов

Общеинститутская кафедра естественно-научных дисциплин
Московский городской педагогический университет
2-й Сельскохозяйственный проезд, 4, Москва, Россия, 129226

В статье анализируется количественная мера информации. Приведены расчеты измерения информации на примере текстов классиков русской поэзии.

Ключевые слова: информация, представление информации, измерение информации, информатика, компьютер.

Слово «информация» произошло от латинского слова īnfōrmatio, что в переводе означает «осведомление, просвещение». Одно из значений указанного слова — «сообщение, осведомляющее о положении дел, о состоянии чего-либо». Понятие «информация» часто объясняется с помощью других понятий, имеющих столь же неопределенное значение, например «сведения», «содержание», «данные» и т.д.

В книге М. Мазура «Количественная теория информации» понятие «информация» определяется следующим образом: «Если наименование некоторого понятия x обозначить T_x, а определяющее его выражение (содержание) через D_x, то информация — это предложение типа T_x есть D_x» [5].

Другими словами, информация — это содержание, заключенное в символе, которым обозначается то или иное понятие как название определенного объекта. Наиболее легко такое толкование рассматриваемого понятия иллюстрируется анализом содержания математических символов. Общеизвестно, что математические знаки (символы) служат для записи математических понятий, предложений и выкладок. Первыми математическими знаками были символы для изображения чисел — цифры, возникновение которых предшествовало письменности. Например, число π — отношение длины окружности к диаметру. В табл. 1 последнее предложение переписано в форме «T_x есть D_x».

Таблица 1

Примеры раскрытия информации

x (понятие)	T_x (наименование понятия)	D_x (выражение понятия)
Число π	π	Отношение длины окружности к диаметру
Функция	$y = f(x)$	Закон по которому произвольному числу x ставится в соответствие строго определенное число y
Производная функции	$f'(x)$	Предел отношения приращения функции к приращению аргумента, если последний стремится к нулю

Здесь же приведены предложения «T_x есть D_x» для раскрытия информации, которая содержится в математических символах $f(x)$ и $f'(x)$.

О роли математических знаков и важности точного определения их смысла русский математик Н.И. Лобачевский писал, что подобно тому, как дар слова обогащает нас мнениями других, так язык математических знаков служит средством еще более совершенным, более точным и ясным, чтобы один передавал другому понятия, которые он приобрел, истину, которую он постигнул, и зависимость между частями, которую он открыл. Но так как мнения могут казаться ложными, оттого что разумеют иначе слова, то всякое суждение в математике останавливается, как скоро перестаем понимать под знаком то, что оно собой представляет [1].

Более того, использование математической символики способствовало созданию математического языка, который преобладает над естественным языком при описании математических знаний.

Выдающийся русский физиолог и психолог И.М. Сеченов при изучении различных форм человеческого мышления определил форму, которую назвал «мышление символами» [6].

Действительно, окружающий мир наполнен различными предметами, которые вместе с их индивидуальными различиями может запоминать человек. Но в силу подмеченного опытным путем закона регистрации впечатлений по сходству в человеческой памяти все сходные предметы, по мнению И.М. Сеченова, смешиваются в так называемые средние итоги.

Эти средние продукты мышления не являются точным воспроизведением действительности, но по смыслу они представляют знаки, заменяющие собой множество однородных предметов.

Такие знаки И.М. Сеченов называл символами первой ступени.

Далее в так называемых средних итогах того или иного предмета человек научился различать отдельные части данного предмета. Расчленение целого предмета на части и оценку математических соотношений между частями И.М. Сеченов назвал символами второй ступени.

Очевидно, что знаковое письмо древнего человека служит доказательством того, что на данном этапе своего умственного развития человечество мыслило символами второй ступени.

При передаче информации в форме сообщения следует отметить, что сообщение — это последовательность символов, набранных из букв некоторого алфа-

вита. Слово — это определенной длины последовательность букв. Слово характеризуется длиной, и длина слов зависит от количества букв алфавита.

В информатике все слова набираются из букв двоичного алфавита $z = (0; 1)$. Для этого существует международный код, с помощью которого буквы русского алфавита кодируются словами из z-алфавита.

Пусть m — длина слова (число букв, число двоичных разрядов) тогда число слов N как число возможных наборов из 0 и 1 длиной m равно

$$N = 2^m. \tag{1}$$

Например, если $m = 1$, то число слов $N = 2$. Это слова 0 и 1. Если же слово $m = 2$, то по (1) будем иметь число $N = 4$. В этом случае это будут слова: 00, 01, 10, 11.

Число слов N — это число сообщений длиной m или информаций, которые можно воспроизвести по каналу связи или в информационном канале.

Для получения количественной меры информации формулу (1) подвергают операции логарифмирования по основанию два, после чего вместо (1) получаем

$$m = \log_2 N. \tag{2}$$

Формула (2) трактуется следующим образом. Если по каналу связи передается слово длиной m, то оно идентифицируется N описательными информациями как максимально возможными. Поэтому считают

$$H = m = \log_2 N. \tag{3}$$

т.е. H — это количество возможных информаций идентифицирующих любое двоичное слово длины m. Другими словами, числом всевозможных двоичных слов диной m можно передать N описательных информаций.

Величина H принимается за количественную меру информации в информатике, а формула (3) была введена Л. Хартли в 1928 г. Таким образом, согласно формуле (3) величина H представляет длину двоичного слова, которая равна количеству двоичных разрядов. Следовательно, единица информации равна одному двоичному разряду, в котором может быть либо 0, либо 1. По этому поводу говорят так: количество информации, заключающееся в одном двоичном разряде, равно одному биту (от англ. binary digit).

Обобщение формулы (3) на случай, когда рассматриваются слова различной длины, выполнено в [1].

Пусть имеет место N слов как объем некоторой статистической выборки, состоящей из нескольких групп слов. Далее предположим, что в пределах каждой группы слова имеют одинаковую длину. Обозначим через n_i объем каждой группы слов, тогда очевидно, что

$$N = \sum n_i.$$

По формуле (2) вычислим длину слова из группы n_1, после будем иметь

$$m_1 = \log_2 n_1. \tag{4}$$

Теперь вычислим разность левых и правых частей в формулах (2) и (4):

$$m - m_1 = \log_2 \left(\frac{N}{n_1} \right) = H_1. \tag{5}$$

Из (5) известно, что H_1 суть длина не идентифицированных слов.

Аналогично можно составить следующие соотношения:

$$H_2 = \log_2\left(\frac{N}{n_2}\right)..., \quad H_k = \log_2\left(\frac{N}{n_k}\right). \tag{6}$$

Затем составим среднюю статистическую сумму

$$H = \frac{1}{N}(n_1H_1 + n_2H_2 + ... + n_kH_k) = \frac{1}{N}\sum n_iH_i,$$

которую с учетом (5), (6) перепишем так:

$$H = \sum\left(\frac{n_i}{N}\right)\log_2\left(\frac{N}{n_i}\right). \tag{7}$$

Если в каждой группе число слов равно единице ($n_i = 1$) и число групп k равно N, то формула (7) переходит в формулу Хартли (3). В противном случае обозначим через p_i частоту появления i-й группы, которую определим общеизвестным способом

$$P_i = \frac{n_i}{N}. \tag{8}$$

Теперь после подстановки (8) в (7) получим формулу Шеннона

$$H = -\sum P_i \log_2 P_i. \tag{9}$$

Очевидно, что формула (9) применима к количественным измерениям информации, содержащейся в текстах естественного языка.

Величину H, определяющуюся по формуле (9), называют энтропией информации.

Группы слов, встречающихся в текстах, можно классифицировать по начальной букве. Тогда величины p_i в (8) будет означать частоты появления слов имеющихся в тексте, на конкретную начальную букву, а энтропию, вычисленную согласно (9) по указанным частотам, будем называть энтропией по начальным буквам и обозначать H_1. Величина H_1 для поэтических текстов Н. Рубцова вычислялась в [2] с помощью программы Microsoft Excel.

Можно также под p_i подрегулировать частоты появления в тексте той или иной буквы русского алфавита. Энтропию, вычисленную по таким частотам, будем обозначать H_2 и называть энтропией текста по всем буквам [3].

В табл. 2 приведены значения энтропий H_1 и H_2, заимствованные из [2; 3], для сорока четырех поэтических текстов Н. Рубцова.

Анализ данных табл. 2 показывает, что диапазон изменения величин H_1 больше диапазона H_2 и есть тексты, для которых числовые значения H_1 и H_2 близки (табл. 3).

Таблица 2

Значения энтропий H_1 и H_2

Стихотворение	H_1 бит	H_2 бит
Элегия	3,6294	4,3116
Ось	4,0043	4,4513
На вокзале	3,9064	4,4802
Весна на берегу Бии	4,0055	4,5382
Прощальная песня	4,0215	4,5462
В лесу	3,1878	4,3346
Ветер всхлипывал словно дитя	3,895	4,3763
У церковных берез	3,9637	4,4996
В московском кремле	7,1349	4,4231
Поэзия	3,4573	4,5084
Сентябрь	3,7171	4,4649
По дороге к морю	8,2939	4,6126
Стоит жара	4,0368	4,4342
Плыть, плыть	6,4871	4,4778
Волнуется море	4,1101	4,5217
Гость молчит и я ни слова	3,7901	4,4911
В пустыне	3,7915	4,4016
Увлекаюсь нечаянно	3,4473	4,2204
В горной деревне	4,0075	4,5289
Мечты	3,7149	4,4679
Видения на холме	4,2156	4,4618
Грани	4,0147	4,1788
По мокрым скверам проходит осень	3,8397	4,2117
В полях смеркалось. Близилась гроза	3,6658	4,1651
Привет Россия	3,9456	3,8839
В горнице	3,8638	4,3709
Родная деревня	4,1048	4,4437
Вологодский пейзаж	4,1059	4,4986
Далекое	7,1822	4,5458
Старик	3,7852	4,4542
Сапоги мои — скрип да скрип	3,8066	4,4202
Памяти матери	3,87	4,3806
В сибирской деревне	4,0113	4,4737
Зимним вечерком	3,9955	4,4054
Меж болотных стволов красовался восток огнеликий	3,9847	4,4701
Синенький платочек		4,1398
Острова свои оберегаем	3,9407	4,4129
А между прочим осень на дворе	4,0818	4,1676
Есть пора — души моей отрада	4,2165	4,4475
Старый конь	3,9524	4,2452
Прекрасное небо голубое		4,4840
На реке Сухоне	3,6389	4,3361
Добрый Филя	3,9008	4,5050
Оттепель	3,7292	4,4821

Таблица 3

Числовые значения H_1 и H_2

Стихотворение	H_1	H_2
Ось	4,0043	4,4513
Весна на берегу Бии	4,0055	4,5382
Прощальная песня	4,0215	4,5462
Стоит жара	4,0368	4,4342
Волнуется море	4,1101	4,5217
Видения на холме	4,2156	4,4618
Грани	4,0147	4,1788
Родная деревня	4,1048	4,4437

Окончание

Стихотворение	H_1	H_2
Вологодский пейзаж	4,1059	4,4986
В сибирской деревне	4,0113	4,4737
А между прочим осень на дворе	4,0818	4,1676
Слез не лей	4,2165	4,4475
Прекрасное небо голубое	4,0479	4,4840

Для более широкого изучения факта близости величин H_1 и H_2 обратимся к текстам классиков русской поэзии. Выдающийся русский поэт Ю. Кузнецов в размышлениях о русской поэзии [7] выделяет в ней две характерные темы. Одна из них — любовная тема, начатая А.С. Пушкиным, а другая — дорожная тема, начатая М.Ю. Лермонтовым.

Таблица 4

Значения H_1 и H_2

Стихотворения	Автор	H_1	H_2
Любовная тема в русской поэзии, начатая А. С. Пушкиным (тема 1)			
Я помню чудное мгновенье	А.С. Пушкин	4,0163	4,4958
Средь шумного бала случайно	А.К. Толстой	3,9922	4,5610
К.Б.	Ф.И. Тютчев	3,2041	4,3764
Сияла ночь	А.А. Фет	4,0213	4,5122
Незнакомка	А.А. Блок	4,0503	4,5093
За дорожной случайной беседой	Ю.П. Кузнецов	4,0407	4,4605
Дорожная тема в русской поэзии, начатая М. Ю. Лермонтовым (тема 2)			
Выхожу один я на дорогу	М.Ю. Лермонтов	4,0871	4,5144
Тройка	Н.А. Некрасов	3,9837	4,5852
Накануне годовщины	Ф.И. Тютчев	3,7152	4,3921
Осенняя воля	А.А. Блок	4,0756	4,5543
Распутье	Ю.П. Кузнецов	3,9345	4,3805

В таблице 4 приведены названия текстов, отобранные Ю. Кузнецовым, и числовые значения H_1 и H_2 для них, заимствованные из [4]. Оказалось, что тексты Н. Рубцова из табл. 3 по числовым значениям H_1 и H_2 близки к текстам табл. 4. Подмеченная близость указанных текстов установлена формально, и только эксперты могут прокомментировать эту близость с литературоведческих позиций.

ЛИТЕРАТУРА

[1] *Бубнов В.А.* О толковании понятия «информация» и о количественной мере информации // Вестник Московского городского педагогического университета. Серия «Естественные науки». — 2009. — № 1. — С. 69—75.

[2] *Бубнов В.А., Ануфриев С.В., Казакова И.С.* Анализ поэтических текстов на уроках литературы с помощью информационных технологий // Информационные технологии в предметной области: Сб. науч. тр. — Вып. 1. / Под ред. В.А. Бубнова. — М.: МГПУ, 2002. — С. 82—102.

[3] *Бубнов В.А., Огородников А.Ю.* Частотный анализ поэтических текстов Н. Рубцова // Информационные технологии в предметной области: Сб. науч. тр. — Вып. 2. / Под ред. В.А. Бубнова. — М.: МГПУ, 2004. — С. 86—111.

[4] *Бубнов В.А., Огородников А.Ф.* Формальный анализ поэтических текстов русской поэзии: Сб. науч. тр. кафедры естественно-научных дисциплин. — Вып. 1. / Под ред. В.А. Бубнова. — М.: МГПУ, 2005. — С. 197—219.
[5] *Мазур М.* Количественная теория информации. — М.: Мир, 1971.
[6] *Сеченов И.М.* Элементы мысли. — СПб.: Питер, 2004.
[7] *Кузнецов Ю.П.* Воззрение // Наш современник. — 2000. — № 1. — С. 101—115.

SYMBOLICAL REPRESENTATION OF THE INFORMATION AND ITS MEASUREMENT

V.A. Bubnov

Chair of natural-science disciplines
The Moscow city pedagogical university
2-nd Selskokhozyajstvennij pr., 4, Moscow, Russia, 129226

In article the quantitative measure of the information is analyzed. Calculations of measurement of the information on an example of texts of classics of Russian poetry are resulted.

Key words: the Information, information representation, information measurement, computer science, the computer.