О КОЛИЧЕСТВЕ ИНФОРМАЦИИ

А. П. Бельтюков, А. Н. Тетерин

Удмуртский государственный университет

Ижевск, Россия

Центральные понятия в математической теории информации – понятия информации и ее количества.

Понятие **информации** – одно из фундаментальных в современной науке. Информацию наряду с веществом и энергией рассматривают в качестве важнейшей сущности мира. Формально определить понятие «информация» чрезвычайно сложно.

В простейшем бытовом понимании с термином «информация» обычно ассоциируются некоторые сведения, данные, знания и т.п. Информация передается в виде **сообщений**, которые и представляют форму существования информации. Примеры сообщений: музыкальное произведение, телепередача, команды регулировщика на перекрестке, распечатанный на принтере текст, результат работы компьютерной программы и т.д. При этом предполагается, что имеются «источник информации» и «получатель информации».

Сообщение от источника к получателю передается посредством какой-нибудь среды, называемой в таком случае «каналом связи». Так, при передаче речевого сообщения в качестве такого канала связи можно рассматривать воздух, в котором распространяются звуковые волны, а в случае передачи письменного сообщения (например, текста, распечатанного на принтере) каналом сообщения можно считать лист бумаги, на котором напечатан текст.

Наше *первое предложение* заключается в том, чтобы при определении информации рассматривать несколько каналов связи. При этом информация, поступающая по одному каналу, может пояснять и уточнять информацию, поступающую по другому каналу.

Наше *второе предложение* заключается в том, чтобы вести неопределенность как естественный и неотъемлемый атрибут информации. Это значит, что значения атрибутов объектов материального мира должны описываться не только с помощью привычных символов (например, 0 и 1 при двоичной кодировке), но и с помощью символа ω – «не определено».

Наше *третье предложение* заключается в том, чтобы учесть процессы кодирования при оценке количества получаемой информации.

При выборке по k символам слов из словаря объема M естественно считать, что мы получаем $-\log_2(D_i/M)$ бит информации, где D_i - количество слов, вырезаемое из словаря последовательностью из k символов. Можно подсчитать количество информации, передаваемое с i-м символом в m-м слове:

$I_{i,m}=\log_2 D_{i,m}/M - \log_2 D_{i-1,m}/M = \log_2 D_{i,m}/D_{i-1,m}$

Тогда естественно получается следующая минимальная оценка информационного объема словаря:

$$I = \sum_{m=1}^{M}\sum_{i=1}^{K_m} I_{i,m}, \qquad (1)$$

где K_m - длина слова в словаре.

Если считать, что энтропия H – средняя длина слов, используемых при двоичном кодировании , то получим, что количество информации I, приходящееся на 1 символ слова из k букв (при двоичном кодировании), определяется так:

$$I = (\log_2 \frac{D_k}{M}) / (\sum_{i=1}^{k} P_i \log_2 \left(\frac{1}{P_i}\right))$$

Среднее количество информации приходящееся на один символ: $I_{cp} = -(\log_2 \frac{1}{M}) / L_{cp}$

L_{cp} – средняя длина слова в словаре.

С учетом первых трех предложений количество информации содержащееся в одном слове сообщения:

$$I_c = I_m + \log_3 rn$$

I_m- может быть получено из (1) с пересчетом по троичному логарифму, г – количество листьев в дереве классификации слов в словаре, n – количество каналов.

Наиболее интересна формула для количества информации в отрезке слова m от символа номер k_i до символа номер l_i в следующей записи:

$$I = \sum_{i=k_m}^{l_m} I_{i,m}.$$

На основе этой формулы малоинформативные символы могут объединяться с соседними символами (по правилам морфологии, чтобы избавиться от лишних вычислений для естественных языков), несущими больше информации, образуя устойчивые сочетания, зависящие от местоположения в слове: приставки, корни, суффиксы, окончания. Причем одни те же сочетания, стоящие в слове на разных местах, получают неодинаковые, но близкие коды, однозначно идентифицирующие их место с использованием символа неопределенности в случае необходимости (что может потребоваться для распознавания слитной речи).

Объектный подход является абстракцией алгоритмического подхода к определению информации. Пусть имеются некоторое множество возможных «событий» U, о которых требуется посылать сообщения. Пусть имеется также некоторый d-символьный алфавит S для формирования этих сообщений в виде цепочек символов на этом алфавите. Пусть

имеется некоторая функция a для интерпретации этих сообщений (S^* - множество всех конечных цепочек в алфавите S):

$$a: S^* \to U.$$

Через $|w|$ будем обозначать число символов (вхождений букв) в слове w. Тогда количеством информации в событии x относительно интерпретации a в объектном смысле (в *d-адической кодировке*) будем считать

$$I_a(x) = \min\{|w|: w \in S^*, a(w) = x\}.$$

При $d = 2$ единица измерения такой информации называется *дит* (***d****it* – ***d***yadic dig***it*** – диадическая цифра). Количество информации в сообщении w относительно интерпретации a в объектном смысле определяется равенством $I_a(w) = I_a(a(w))$. Основной недостаток такого подхода – неэффективность: значение

$$\min\{|w|: w \in S^*, a(w) = x\}$$

иногда очень трудно (если вообще возможно) вычислить. Один из способов обойти этот недостаток – введение кодирующей функции b:

$$b: U \to S^*$$

со следующим свойством: $a(b(x)) = x$ для любого x из U. Тогда количество информации в объекте x относительно интерпретации a и кодировки b определяется равенством

$$I_{a,b}(x) = |b(x)|.$$

Количество информации в сообщении w относительно интерпретации a и кодировки b определяется равенством $I_{a,b}(w) = I_{a,b}(a(w))$. Основной недостаток такого варианта определения количества информации – чрезмерная зависимость от кодировки.

Другой общий недостаток объектного подхода – неаддитивность: количество информации, переданное двумя сообщениями, вообще говоря, не равно сумме количеств информации в этих сообщениях, а может содержать, например, ещё и логарифмическую добавку, так как длина сообщения также содержит некоторую информацию.