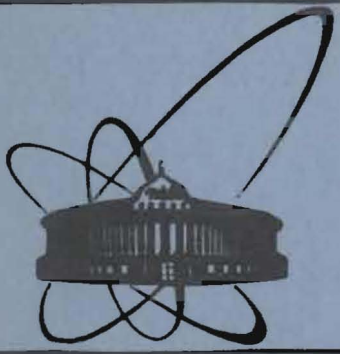

объединенный
институт
ядерных
исследований
дубна

Р19-85-192

В.И.Корогодин, Ч.Файси

КОЛИЧЕСТВО ИНФОРМАЦИИ И ЕМКОСТЬ "ИНФОРМАЦИОННОЙ ТАРЫ"

Направлено в "International Journal of Systems Science"

1985

1.

В[1,2] было предложено различать понятия "Информация" и "Информационная тара". Под информацией понималась содержательная сторона сообщений, а под информационной тарой - сигналы, символы или буквы, совокупности которых составляют сообщения. "Сообщением" в дальнейшем изложении мы будем называть любую конечную последовательность знаков, букв или символов.

Очевидно, что для людей сообщения представляют ценность лишь постольку, поскольку они осмысленны, т.е. содержат информацию. Технические системы связи были созданы для обмена именно такими сообщениями. Существенно, что практически все системы связи пригодны для передачи и/или приема сообщений любого содержания. Эта универсальность систем связи, отражающая инвариантность информации по отношению к ее носителям[1,3], возможна потому, что любую информацию можно "записать" на любом языке, а знакам этого языка можно поставить в соответствие отдельные сигналы, передаваемые по каналам связи.

Так как информация "фиксируется" в сообщении последовательностью составляющих его сигналов или букв, то последовательность эта для сообщений, содержащих какую-либо информацию, должна быть строго определенной. Сообщения, представляющие собой случайные последовательности букв, информации не содержат. Таким образом, хотя информацию можно передавать только в форме сообщений, сами сообщения могут либо содержать, либо не содержать информацию: информационная тара может быть либо "заполненной", либо оставаться "пустой". Это обстоятельство отмечали еще Шеннон[4] и Кастлер[5], но поскольку их интересовала передача только сообщений, они не пошли дальше простой констатации факта.

2.

Изложенные выше соображения заставляют по-иному трактовать содержание понятия "Количество информации", нежели это было принято ранее, и позволяют выявить такие аспекты шенноновской концепции, которым до сих пор не уделяли должного внимания. Однако прежде чем приступить к обсуждению этих вопросов, следует коротко напомнить основные положения классической теории информации.

Как известно, основы математической теории связи, получившей в последующем название "Теория информации", были сформулированы Шенноном [4] в ходе работы по оптимизации технических систем связи.

Работа по оптимизации технических систем связи требовала, прежде всего, умения определять "пропускную способность" каналов связи, т.е. то количество информации, которое может быть передано по данному каналу за единицу времени. Но по каналам связи непосредственно передают не информацию, а сообщения, состоящие из совокупностей сигналов и букв. Поэтому, прежде чем приступить к определению пропускной способности каналов связи, следовало установить, какое количество информации может быть передано с помощью одной буквы того или иного алфавита.

Если последовательные буквы сообщения выбираются независимо, с фиксированными вероятностями p_i, то подходящей мерой количества информации, приходящегося на одну букву, оказывается величина

$$H = -k \sum_{i=1}^{n} p_i \log_q p_i , \qquad /1/$$

которую Шеннон и назвал "количеством информации", или "энтропией источника" /из-за формального сходства уравнения /1/ с определением термодинамической энтропии/. Здесь p_i - частота встречаемости i-й буквы в языке, на котором составлено сообщение /$i = 1,2,\ldots,n$/; q - основание логарифмов; k - коэффициент пропорциональности, величина которого зависит от q и от избранных единиц измерения количества информации; знак "минус" перед k поставлен для того, чтобы величина H всегда оставалась положительной.

Шенноновское количество информации можно интерпретировать как взвешенную среднюю того количества информации, которое несет i-я буква:

$$H_i = -k \log_q p_i . \qquad /2/$$

Тогда суммарное количество информации для сообщения, состоящего из M букв, будет

$$H_M = -k \sum_{i=1}^{n} m_i \log_q p_i , \qquad /3/$$

где m_i - число i-х букв в сообщении ($M = \sum_{i=1}^{n} m_i$). Шеннон показал, что с увеличением длины сообщения M почти все сообщения будут иметь "типичный состав": $m_i/M \to p_i$. Следовательно,

$$H_{M\to\infty} \to -k M \sum_{i=1}^{n} p_i \log_q p_i . \qquad /4/$$

Это и есть сущность шенноновской концепции - тот фундамент, на котором построена классическая теория информации /эта теория, естественно, была обобщена и для языков с более сложной статистической структурой, а также для непрерывных сигналов, но для настоящего рассмотрения достаточно и вышеизложенного/.

3.

Обратим теперь внимание на то, что в шенноновской концепции понятие "Информация" имеет два смысла. Действительно, шенноновская концепция явно не отвергает общепринятого представления об информации как о содержательной стороне сообщений. В то же время формулы /1/ и /3/ не учитывают порядка следования букв в сообщении, благодаря чему оказывается, что случайная комбинация M букв якобы содержит такое же количество информации, что и составленный из них осмысленный текст. Но случайная последовательность букв не может содержать такой же информации, что и осмысленное сообщение! Тогда о какой информации идет речь в шенноновской концепции?

Согласно /2/, величина H_i возрастает с уменьшением частоты p_i встречаемости i-й буквы в языке. Некоторые авторы /см., напр., [5-7] / склонны были интерпретировать это так, как будто бы адресат, воспринимающий очередную букву сообщения, "получает" тем больше информации, чем реже эта буква встречается в данном языке, независимо от ее места в ряду составляющих сообщение букв. Такая точка зрения, однако, не разрешала, а усугубляла отмеченное выше противоречие: она предполагала, что восприятие адресатом такой информации не имеет ничего общего с его восприятием содержательной стороны сообщения. Тогда возникает вопрос: зачем измерять количество той информации, которая не имеет ничего общего с информацией, содержащейся в сообщении, для обмена которой и были созданы технические системы связи? Более того, возникает недоумение: каким образом эта процедура, как будто лишенная особого смысла, могла привести основанную на ней теорию информации к ряду достижений в разных прикладных областях [8]?

Прежде чем разобраться в этом, посмотрим, к каким следствиям приводит принятие определения количества информации, содержащегося в формуле /2/, и целесообразно ли вообще величину H называть "количеством информации".

4.

Заметим, что формула /2/ содержит в себе соблазн распространить даваемое ею определение понятия "Количество информации" за пределы математической теории связи.

Действительно, если такое событие, как появление очередной буквы сообщения, "несет с собой" информацию, количество которой определяется только его частотной характеристикой p_i, то почему бы не приписать информационное содержание осуществлению любого события или появлению любого объекта, независимо от их природы? Если частота осуществления какого-либо события есть P, то почему бы не считать, что его осуществление "несет с собой" $H = -k \log_q P$ информации? При этом появление очередной буквы сообщения, передаваемого по каналу связи, можно рассматривать как частный случай такой, более общей, информационной концепции.

Подобные обобщения и были сделаны вскоре после публикации работы Шеннона /4/, причем их авторы /5-7/ особым достоинством сформулированных представлений считали именно независимость такого способа оценки количества информации от природы "несущего ее" события или объекта. Эта точка зрения сохраняется в теории информации до сих пор /9,10/ .

Идея разделения количества информации и ее смыслового содержания, или семантики сообщения, безусловно, правильная. Однако эта идея отнюдь не отрицает наличия у информации семантики, а лишь предполагает возможность независимого рассмотрения ее количественного и семантического аспектов /11/. Когда же мы определяем количество информации как H-функцию частотной характеристики любого события, независимо от того, является ли оно появлением очередной буквы при передаче сообщения или вспышкой молнии, то вопрос о семантике информации вообще исчезает. Естественно спросить себя: насколько эвристичен и правомочен такой подход?

Действительно, при таком определении количества информации теряет смысл не только вопрос о ее семантике, но и другие обычно связанные со словом "Информация" вопросы: "от кого она поступила?", для кого предназначена? и т.п. Информация, количество которой определяется таким образом, утрачивает все свои особенности и свойства, в том числе и свойство инвариантности по отношению к природе носителя: теперь каждый "носитель" несет с собой только "свою" информацию, отражающую только его частотную характеристику. При этом, чтобы "получить" такую информацию, следует заранее знать частотную характеристику соответствующего события!

Такой подход приводит также к выводу о том, что информация "содержится" во всех без исключения событиях и объектах окружающего нас мира, являясь чем-то вроде третьей ипостаси материи, наряду с массой и энергией; что информация не возникает и не уничтожается, а лишь "передается" от одних тел другим. При последовательном проведении этот подход приводит к так называемому "негэнтропийному принципу информации", отождествляющему информацию с энтропией, а коэффициент k формул /2/-/4/ - с константой Больцмана /7,10/ ; бессодержательность подобных построений неоднократно отмечалась в литературе /см., напр., /1,12,13/ /.

Сказанного, по-видимому, достаточно, чтобы показать, что распространение шенноновского определения количества информации за пределы математической теории связи не имело позитивных последствий, кроме, разве что, более тщательной отработки формального аппарата.

5.

Как мы помним, необычайные "приключения", которые довелось испытать понятию "Информация", начались с того, что Шеннон дал определение количества информации, не зависящее от ее семантики. Чтобы освободиться от семантической специфики сообщений, он использовал в своих построениях не реальные сообщения, а их модели - случайные последовательности букв. Но модель сообщения не просто лишена семантики - она вообще не может содержать информацию! Поэтому, вводя определение количества информации на примере таких моделей, следовало бы подчеркнуть, что это - количество той информации, которую могло бы содержать сообщение, состоящее из M букв, если бы оно представляло собой не случайную их последовательность /модель!/, а осмысленный текст /14/. Такая оговорка не повлияла бы на развитие математической теории связи - ведь по каналам связи передают последовательности букв, а не информацию. Это, однако, весьма благотворно сказалось бы на последующей разработке логических основ теории информации, избавив ее от многолетних блужданий по лабиринтам двусмысленных и нечетких определений.

Действительно, если бы с самого начала было ясно, что формула /2/ отражает не то количество информации, которое якобы связано с появлением i-й буквы при передаче сообщения, а то ее количество, которое может приходиться на i-ю букву осмысленного текста, то это могло бы иметь минимум два последствия. Во-первых, оно ограничило бы область приложения формулы /2/ кругом объектов, представляющих собой, хотя бы потенциально, сообщения или их компоненты - отдельные буквы. Во-вторых, было бы ясно, что H_i есть оценка не количества самой информации, а лишь "информационной емкости" тех объектов, которые служат или могут служить ее носителями, - т.е. могут выступать в роли букв какого-либо алфавита того или иного языка. Именно поэтому величина H_i не зависит от положения i-й буквы в сообщении - ведь информационная емкость буквы не может зависеть от того, рассматривается ли она изолированно, или как компонент случайной последовательности, или как буква осмысленного текста. Так, емкость бочонка не изменится, будет ли он пуст или заполнен водой, уксусом или вином. Все это, конечно, предотвратило бы возникновение той двусмысленности в интерпретации понятия "Информация", о которой шла речь выше.

В каких единицах логичнее всего измерять емкость информационной тары? Чтобы ответить на этот вопрос, обратимся к формуле /1/. Можно сразу заметить, что числовой коэффициент k здесь излишний: при переходе к новому основанию логарифмов $q' = q^{1/k}$, H будет выражаться как

$$H = -\sum_{i=1}^{n} p_i \log_{q'} p_i .$$

При этом значение H не меняется, но коэффициент k устраняется.

В симметричном случае равных вероятностей, $p_i = \frac{1}{n}$,

$$H = \log_{q'} n .$$

H будет равно единице при $q' = n$. Если принять $q' = 2$, то формула /1/ приобретает вид

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i . \qquad /5/$$

Соответственно H=1 при n=2 и $p_1 = p_2 = 0{,}5$. Такую емкость элементарного носителя информации, согласно традиции, будем называть "двоичной единицей" или битом.

Таким образом, при выражении информационной емкости буквы или сообщения в битах и использовании основания логарифмов q=2, коэффициент k в формулах /1/-/4/ становится ненужным. Если вспомнить о том, как жонглировали этим коэффициентом Бриллюэн и другие адепты негэнтропийного принципа информации, то невольно придешь к выводу, что изгнать этот коэффициент из математического аппарата теории информации следовало уже давно.

Заметим, что бит имеет реальный физический смысл: это емкость одной буквы кода, наиболее компактного из всех возможных кодов с бинарным алфавитом, когда $p_1 = p_2 = 0{,}5$. Только для такого кода при выражении в битах H = M . Выражение в битах информационной емкости какого-либо сообщения означает, что она равна информационной емкости H букв такого кода.

Физический смысл бита как единицы измерения емкости информационной тары, можно также интерпретировать как "вместимость" таких элементарных носителей информации, которые соответствуют одному дихотомическому выбору, т.е. могут находиться в одном из двух равновероятных состояний. Из этого следует, что /при равновероятных состояниях/ максимальная емкость любого носителя информации равна двоичному логарифму числа возможных его состояний. Анализ связи информационной емкости носителей с их физическими особенностями дан в /15/ .

Теперь рассмотрим вопрос об определении количества самой информации.

Если сообщения интерпретировать как информационную тару, то количество содержащейся в них информации естественно выражать в объеме "заполненной" ею тары. Как, однако, можно оценить величину этого объема?

Представить себе это можно следующим образом. Пусть дано некоторое сообщение, содержащее данную информацию. Если его уплотнять, вычеркивая повторы, заменяя пространные обороты более лаконичными и т.д., то в конце концов можно приблизиться к максимально компактному тексту, в котором замена, выпадение или вставка даже одной буквы либо исказит смысл, либо увеличит информационную емкость. Будем считать, что в максимально-компактном тексте информационная тара полностью загружена информацией. Емкость такого максимально-компактного сообщения, содержащего данную информацию, и можно считать мерой ее количества.

Очевидно, что при таком определении количества информации его следует выражать в тех же единицах, что и емкость информационной тары, т.е. в битах, подобно тому, как в литрах выражают и емкость сосудов для вина, и количество вина, в них содержащегося. Для обозначения емкости информационной тары сохраним символ H, а количество самой информации будем обозначать символом B.

Хотя H и B можно выражать в одних и тех же единицах измерения, по своей величине H и B, относящиеся к одному и тому же сообщению, могут существенно различаться между собой.

Действительно, пусть имеется максимально компактное сообщение из M букв какого-либо алфавита. Емкость тары такого сообщения, согласно /4/ и /5/, равна

$$H_M \approx -M \sum_{i=1}^{n} p_i \log_2 p_i \text{ бит}. \qquad /6/$$

Поскольку этот текст максимально компактен, то $B_M = H_M$. Перетасовывая буквы этого сообщения случайным образом, получаем случайную последовательность, которая не содержит никакой информации, т.е. $B_M = 0$, хотя H_M осталось неизменным. Следовательно, в реальных сообщениях, которые могут быть в разной мере загружены информацией, величина B заключена между границами

$$0 \le B \le H . \qquad /7/$$

Таким образом, количество информации B, содержащейся в реальных сообщениях, не может превышать их информационную емкость H.

Как мы уже отмечали, согласно свойству инвариантности, одна и та же информация может быть "записана" с помощью разных алфавитов и кодов. Поэтому, когда мы выражаем количество информа-

ции в битах, это, по существу, означает, что для фиксации B бит информации с помощью бинарного кода требуется не менее B букв. Таким образом, количество бит информации - это число букв бинарного кода, необходимое и достаточное для ее максимально компактной записи.

8.

Разные количественные аспекты информации по-разному проявляются в разных ситуациях.

Прежде всего, следует подчеркнуть, что шенноновская концепция количества информации, которую логичнее интерпретировать как концепцию емкости информационной тары, легла в основу математической теории связи и получила широкое практическое применение не случайно; разделение понятий "Информация" и "Информационная тара" не может повлиять на эту ситуацию. Тому есть два веских основания. Во-первых, по каналам связи передаются сообщения, а не "голая" информация. Информационная же емкость отдельных букв и составленных из них сообщений, согласно определению, обусловливается статистическими характеристиками алфавитов, кодов и языков, что и позволяет решать задачи, связанные с оптимизацией способов передачи сообщений, повышением их помехоустойчивости и переводом с одних систем записи на другие. Во-вторых, только информационная емкость сообщений поддается простой оценке, тогда как количество самой информации можно оценить только сверху, и то после специального исследования возможно более компактных записей.

Если отдельные буквы какого-либо алфавита рассматривать как элементарные носители информации, то из формулы /2/ следует, что их информационная емкость тем больше, чем реже они встречаются в данном языке, а средняя инфорционная емкость одной буквы обычно тем меньше, чем меньше букв в данном алфавите. При перекодировании сообщения в алфавит с меньшим числом букв длина записи возрастает. Все это наводит на мысль о том, что инвариантность информации может означать не только возможность записи любой информации на любом носителе и на любом языке, но и то, что одна и та же информация при записи ее на разных языках должна выражаться в максимально компактных сообщениях, обладающих примерно одинаковой информационной емкостью. Информационная емкость сообщений должна оставаться примерно одинаковой и при переводе любых осмысленных текстов с одного языка на другой. Если определение информационной емкости, задаваемое вероятностным подходом, действительно отражает истинное положение вещей, то с помощью формулы /6/ можно оценивать длину M текстов на любом языке и при любой заданной системе кодирования, достаточную для записи на этом языке любых сообщений,

составленных на любых других языках. При использовании же бинарного кода размер сообщения, равный числу M составляющих его букв, при достаточно большом их количестве не должен зависеть от того, на каком языке оно составлено. Это - весьма нетривиальное утверждение, характеризующее языки в терминах в теории информации, в память автора математической теории связи можно назвать "Правилом Шеннона".

Обратим теперь внимание на то, что представление о количестве информации B, содержащейся в сообщении, как об информационной емкости максимально компактного текста, пригодного для ее записи, сближает использованный в нашей работе вероятностный подход к определению этого понятия с алгоритмическим подходом, предложенным Колмогоровым [16,17]. Согласно [16,17], информацию о каком-либо объекте можно представить как записанный бинарным кодом алгоритм минимальной длины, пригодный для построения этого объекта; число букв такой записи и будет равно числу битов этой информации. Однако, как уже отмечалось [2], в виде алгоритма, пригодного для построения какого-либо "оператора", можно представить любую информацию, рассматриваемую как содержательную сторону сообщения. Следовательно, число битов такой информации, равное числу букв бинарного кода, которое необходимо и достаточно для ее максимально-компактной записи, и есть то определение ее количества, к которому приводит алгоритмический подход.

Важно отметить, что количество информации B однозначно определяется семантикой, ее смысловым содержанием, в то время как емкость тары H может быть сколь угодно большой /но не меньше B/.

Наконец, остановимся коротко на вопросе, который возникает при сопоставлении информационной емкости сообщений H и количества содержащейся в них информации B. Мы уже отмечали, что в общем случае B всегда меньше H. Чем же "заполнена" разность между H и B, играет ли она какую-либо роль в обмене информацией и какую именно? Можно предположить, что та часть сообщения, которая представлена разностью $H-B$, выполняет в обмене информацией по меньшей мере три функции. Во-первых, это функция обеспечения помехоустойчивости за счет шенноновской "избыточности" сообщений [4]. Во-вторых, это "адаптационная функция", адаптирующая сообщения к грамматическим требованиям тех языков, на которых они составлены /а также к другим конвенциям общения, напр., к принятой форме заявления или официального письма/, и тем самым как бы способствующая проникновению информации сквозь "фильтры", присущие адресатам [18]. В-третьих, это "эмоциональная функция", служащая для настраивания адресата на восприятие данного сообщения и на то или иное отношение к содержащейся в нем информации [19]. Более детальное рассмотрение этих вопросов, однако, выходит за рамки задач данной работы.

Авторы благодарят проф. М.И.Подгорецкого за обсуждение работы в рукописи и конструктивные замечания.

ЛИТЕРАТУРА

1. Серавин Л.Н. Теория информации с точки зрения биолога. Изд. ЛГУ, Л., 1973.
2. Корогодин В.И. Биофизика, 1983, 28, 1, с.171-178.
3. Дубровский Д.И. Информация, сознание, мозг. "Высшая школа", М., 1980.
4. Шеннон К. В кн.: Работы по теории информации и кибернетике. ИЛЛ., М., 1964, с. 243.
5. Кастлер Г. В кн.: Теория информации в биологии. ИИЛ, М., 1960, с.9-53.
6. Эшби У.Р. Введение в кибернетику. ИИЛ, М., 1959.
7. Бриллюэн Л. Наука и теория информации. Гос. изд-во физ.-мат.лит., М., 1960.
8. Коган И.М. Прикладная теория информации. "Радио и связь", М., 1981.
9. Стратанович Р.Л. Теория информации. "Советское радио", М., 1975.
10. Волькенштейн М.В. Биофизика. "Наука", М., 1981.
11. Харкевич А.А. Проблемы кибернетики. Физматгиз, М., 1960, вып. 4, с.53.
12. Оксак А.И. Философские науки, 1972, №5, с.68.
13. Сольен Дж. Кодирование сенсорной информации в нервной системе млекопитающих. ИИЛ, М., 1975.
14. Харкевич А.А. Очерки общей теории связи. Гостехтеоретиздат., М., 1955.
15. Романовский Ю.М., Степанова Н.В., Чернавский Д.С. Математическая биофизика. "Наука", М., 1984.
16. Колмогоров А.Н. Пробл. передачи информации, 1965, 1, №1, с.3-11.
17. Колмогоров А.Н. Пробл. передачи информации, 1969, 5, №3, с.3-7.
18. Налимов В.В. Вероятностная модель языка. "Наука", М., 1979.
19. Фейнберг Е.Л. Кибернетика, логика, искусство. "Радио и связь", М., 1981.

Рукопись поступила в издательский отдел
18 марта 1985 года.

Вниманию организаций и лиц, заинтересованных в получении публикаций Объединенного института ядерных исследований

Принимается подписка на препринты и сообщения Объединенного института ядерных исследований.

Установлена следующая стоимость подписки на 12 месяцев на издания ОИЯИ, включая пересылку, по отдельным тематическим категориям:

ИНДЕКС	ТЕМАТИКА	Цена подписки на год
1.	Экспериментальная физика высоких энергий	10 р. 80 коп.
2.	Теоретическая физика высоких энергий	17 р. 80 коп.
3.	Экспериментальная нейтронная физика	4 р. 80 коп.
4.	Теоретическая физика низких энергий	8 р. 80 коп.
5.	Математика	4 р. 80 коп.
6.	Ядерная спектроскопия и радиохимия	4 р. 80 коп.
7.	Физика тяжелых ионов	2 р. 85 коп.
8.	Криогеника	2 р. 85 коп.
9.	Ускорители	7 р. 80 коп.
10.	Автоматизация обработки экспериментальных данных	7 р. 80 коп.
11.	Вычислительная математика и техника	6 р. 80 коп.
12.	Химия	1 р. 70 коп.
13.	Техника физического эксперимента	8 р. 80 коп.
14.	Исследования твердых тел и жидкостей ядерными методами	1 р. 70 коп.
15.	Экспериментальная физика ядерных реакций при низких энергиях	1 р. 50 коп.
16.	Дозиметрия и физика защиты	1 р. 90 коп.
17.	Теория конденсированного состояния	6 р. 80 коп.
18.	Использование результатов и методов фундаментальных физических исследований в смежных областях науки и техники	2 р. 35 коп.
19.	Биофизика	1 р. 20 коп.

Подписка может быть оформлена с любого месяца текущего года.

По всем вопросам оформления подписки следует обращаться в издательский отдел ОИЯИ по адресу: 101000 Москва, Главпочтампт, п/я 79.