6. Лекция: Меры информации в системе

Рассматриваются различные способы введения меры измерения количества информации, их положительные и отрицательные стороны, связь с изменением информации в системе, примеры.

Цель лекции: введение в различные способы задания мер для измерения количества информации, их критический сравнительный анализ, основные связи информации и энтропии системы.

В предыдущей лекции было отмечено, что информация может пониматься и интерпретироваться в различных проблемах, предметных областях по-разному. Вследствие этого, имеются различные подходы к определению измерения информации и различные способы введения меры *количества информации*.

Количество информации - числовая величина, адекватно характеризующая актуализируемую информацию по разнообразию, сложности, структурированности (упорядоченности), определенности, выбору состояний отображаемой системы.

Если рассматривается некоторая система, которая может принимать одно из n возможных состояний, то актуальной задачей является задача оценки этого выбора, исхода. Такой оценкой может стать мера информации (события).

Мера, как было сказано выше, - непрерывная действительная неотрицательная функция, определенная на множестве событий и являющаяся аддитивной (мера суммы равна сумме мер).

Меры могут быть статические и динамические, в зависимости от того, какую информацию они позволяют оценивать: статическую (не актуализированную; на самом деле оцениваются сообщения без учета ресурсов и формы актуализации) или динамическую (актуализированную т.е. оцениваются также и затраты ресурсов для актуализации информации).

Ниже мы не всегда будем, в основном, для большей убедительности и большего содержательного понимания, проводить четкие математические границы между понятиями "*количество информации*" и "мера *количества информации*", но строгому читателю необходимо все время задавать достаточно важные вопросы: о *количестве информации* или о мере информации в конкретной последовательности событий идет речь? о детерминированной или стохастической информации идет речь? какова мера измерения *количества информации* и насколько она адекватна?

1. ***Мера Р. Хартли***. Пусть имеется `N` состояний системы `S` или `N` опытов с различными, равновозможными, последовательными состояниями системы. Если каждое состояние системы закодировать, например, двоичными кодами определенной длины `d`, то эту длину необходимо выбрать так, чтобы число всех различных комбинаций было бы

не меньше, чем N. Наименьшее число, при котором это возможно, называется мерой разнообразия множества состояний системы и задается формулой Р. Хартли: $H=k\log_a N$, где k - коэффициент пропорциональности (масштабирования, в зависимости от выбранной единицы измерения меры), a - основание системы меры.

Если измерение ведется в экспоненциальной системе, то $k=1$, $H=\ln N$ (нат); если измерение было произведено в двоичной системе, то $k=1/\ln 2$, $H=\log_2 N$ (бит); если измерение было произведено в десятичной системе, то $k=1/\ln 10$, $H=\lg N$ (дит).

Пример. Чтобы узнать положение точки в системе из двух клеток т.е. получить некоторую информацию, необходимо задать 1 вопрос ("Левая или правая клетка?"). Узнав положение точки, мы увеличиваем суммарную информацию о системе на 1 бит ($I=\log_2 2$). Для системы из четырех клеток необходимо задать 2 аналогичных вопроса, а информация равна 2 битам ($I=\log_2 4$). Если же система имеет n различных состояний, то максимальное *количество информации* будет определяться по формуле: $I=\log_2 n$.

Справедливо утверждение Хартли: если в некотором множестве $X=\{x_1, x_2, \ldots, x_n\}$ необходимо выделить произвольный элемент $x_i \in X$, то для того, чтобы выделить (найти) его, необходимо получить не менее $\log_a n$ (единиц) информации.

Если N - число возможных равновероятных исходов, то величина $k\ln N$ представляет собой меру нашего незнания о системе.

По Хартли, для того, чтобы мера информации имела практическую ценность, она должна быть такова, чтобы отражать *количество информации* пропорционально числу выборов.

Пример. Имеются 192 монеты. Известно, что одна из них - фальшивая, например, более легкая по весу. Определим, сколько взвешиваний нужно произвести, чтобы выявить ее. Если положить на весы равное количество монет, то получим 3 независимые возможности: а) левая чашка ниже; б) правая чашка ниже; в) чашки уравновешены. Таким образом, каждое взвешивание дает *количество информации* $I=\log_2 3$, следовательно, для определения фальшивой монеты нужно сделать не менее k взвешиваний, где наименьшее k удовлетворяет условию $\log_2 3^k \geq \log_2 192$. Отсюда, $k \geq 5$ или, $k=4$ (или $k=5$ - если считать за одно взвешивание и последнее, очевидное для определения монеты). Итак, необходимо сделать не менее 5 взвешиваний (достаточно 5).

Пример. ДНК человека можно представить себе как некоторое слово в четырехбуквенном алфавите, где каждой буквой помечается звено цепи ДНК или нуклеотид. Определим, сколько информации (в битах) содержит ДНК, если в нем содержится примерно $1,5\times 10^{23}$ нуклеотидов (есть и другие оценки этого объема, но мы рассмотрим данный вариант). На один нуклеотид приходится $\log_2(4)=2$ (бит)

информации. Следовательно, структура ДНК в организме человека позволяет хранить 3×10^{23} бит информации. Это вся информация, сюда входит и избыточная. Реально используемой - структурированной в памяти человека информации, - гораздо меньше. В связи с этим, заметим, что человек за среднюю продолжительность жизни использует около 5-6% нейронов (нервных клеток мозга - "ячеек ОЗУ человека"). Генетический код - чрезвычайно сложная и упорядоченная система записи информации. Информация, заложенная в генетическом коде (по учению Дарвина), накапливалась многие тысячелетия. Хромосомные структуры - своеобразный шифровальный код, при клеточном делении создаются копии шифра, каждая хромосома - удваивается, в каждой клетке имеется шифровальный код, при этом каждый человек получает, как правило, свой набор хромосом (код) от матери и от отца. Шифровальный код разворачивает процесс эволюции человека. Вся жизнь, как отмечал Э. Шредингер, "упорядоченное и закономерное поведение материи, основанное ... на существовании упорядоченности, которая поддерживается все время".

Формула Хартли отвлечена от семантических и качественных, индивидуальных свойств рассматриваемой системы (качества информации в проявлениях системы с помощью рассматриваемых N состояний системы). Это основная и положительная сторона формулы. Но имеется основная и отрицательная ее сторона: формула не учитывает различимость и различность рассматриваемых N состояний системы.

Уменьшение (увеличение) H может свидетельствовать об уменьшении (увеличении) разнообразия состояний N системы. Обратное, как это следует из формулы Хартли (так как основание логарифма больше 1!), - также верно.

2. ***Мера К. Шеннона***. Формула Шеннона дает оценку информации независимо, отвлеченно от ее смысла:

$$I = -\sum_{i=1}^{n} p_i \log_2 p_i$$

где n - число состояний системы; p_i - вероятность (или относительная частота) перехода системы в i-е состояние, причем сумма всех p_i равна 1.

Если все состояния равновероятны (т.е. $p_i=1/n$), то $I=\log_2 n$.

К. Шенноном доказана теорема о единственности меры *количества информации*. Для случая равномерного закона распределения плотности вероятности *мера Шеннона* совпадает с *мерой Хартли*. Справедливость и достаточная универсальность формул Хартли и Шеннона подтверждается и данными нейропсихологии.

Пример. Время t реакции испытуемого на выбор предмета из имеющихся N предметов линейно зависит от $\log_2 N$: $t=200+180\log_2 N$ (мс). По аналогичному закону

изменяется и время передачи информации в живом организме. Один из опытов по определению психофизиологических реакций человека состоял в том, что перед испытуемым большое количество раз зажигалась одна из n лампочек, на которую он должен был указать в ходе эксперимента. Оказалось, что среднее время, необходимое для правильного ответа испытуемого, пропорционально не числу n лампочек, а именно величине I, определяемой по формуле Шеннона, где p_i - вероятность зажечь лампочку номер i

Легко видеть, что в общем случае

$$I = -\sum_{i=1}^{n} p_i \log_2 p_i \le \log_2 n$$

Если выбор i-го варианта предопределен заранее (выбора, собственно говоря, нет, $p_i=1$), то $I=0$.

Сообщение о наступлении события с меньшей вероятностью несет в себе больше информации, чем сообщение о наступлении события с большей вероятностью. Сообщение о наступлении достоверно наступающего события несет в себе нулевую информацию (и это вполне ясно: событие всё равно произойдет когда-либо).

Пример. Если положение точки в системе известно, в частности, она - в k-ой клетке, т.е. все $p_i=0$, кроме $p_k=1$, то тогда $I=\log_2 1=0$ и мы здесь новой информации не получаем (как и следовало ожидать).

Пример. Выясним, сколько бит информации несет произвольное двузначное число со всеми значащими цифрами (отвлекаясь при этом от его конкретного числового значения, т.е. каждая из возможных цифр может появиться на данном месте, в данном разряде с одинаковой вероятностью). Так как таких чисел может быть всего 90 (10-99), то информации будет количество $I=\log_2 90$ или приблизительно $I=6,5$. Так как в таких числах значащая первая цифра имеет 9 значений (1-9), а вторая - 10 значений (0-9), то $I=\log_2 90=\log_2 9+\log_2 10$. Приблизительное значение $\log_2 10$ равно 3,32. Итак, сообщение в одну десятичную единицу несет в себе в 3,32 больше информации, чем в одну двоичную единицу (чем $\log_2 2=1$), а вторая цифра, например, в числе аа, несет в себе больше информации, чем первая (если цифры а обоих разрядов неизвестны; если же эти цифры а известны, то выбора нет и информация равна нулю).

Если в формуле Шеннона обозначить $f_i=-n\log_2 p_i$, то получим, что I можно понимать как среднеарифметическое величин f_i.

Отсюда, f_i можно интерпретировать как информационное содержание символа алфавита с индексом i и величиной p_i вероятности появления этого символа в сообщении, передающем информацию.

Пример. Пусть рассматривается алфавит из двух символов русского языка - "к" и "а". Относительные частоты встречаемости этих букв в частотном словаре русского языка равны соответственно $p_1=0.028$, $p_2=0.062$. Возьмем произвольное слово p длины N из k букв "к" и $m\ (k+m=N)$ букв "а" над этим алфавитом. Число всех таких возможных слов, как это следует из комбинаторики, равно $n=N!/(k!\ m!)$. Оценим *количество информации* в таком слове: $I=\log_2 n=\ln n/\ln 2=\log_2 e[\ln N!-\ln k!-\ln m!]$. Используя известную формулу Стирлинга (эта формула, как известно из математического анализа, достаточно точна при больших N, например, при $N>100$) - $N!\approx(N/e)^N$), а точнее, ее важное следствие, - $\ln N!\approx N(\ln N-1)$), получаем оценку *количества информации* (в битах) на 1 символ любого слова:

$$I_1=I/N\approx(\log_2 e/N)[(k+m)(\ln N-1)-k(\ln k-1)-m(\ln m-1)]=$$
$$=(\log_2 e/N)[k\ln(N/k)-m\ln(N/m)]=$$
$$=-\log_2 e[(k/N)\ln(k/N)+(m/N)\ln(m/N)]\le$$
$$\le -\log_2 e\,[p_1\ln p_1+p_2\ln p_2]=$$
$$=-\log_2 e[0{,}028\ \ln 0{,}028+0{,}062\ \ln 0{,}062]\approx 0{,}235.$$

Пример. В сообщении 4 буквы "а", 2 буквы "б", 1 буква "и", 6 букв "р". Определим *количество информации* в одном таком (из всех возможных) сообщений. Число N различных сообщений длиной 13 букв будет равно величине: $N=13!/(4!\times 2!\times 1!\times 6!)=180180$. *Количество информации* I в одном сообщении будет равно величине: $I=\log_2(N)=\log_2 180180\approx 18$ (бит).

Если k - коэффициент Больцмана, известный в физике как $k=1.38\times 10^{-16}$ эрг/град, то выражение

$$S=-k\sum_{i=1}^{n} p_i \ln p_i$$

в термодинамике известно как *энтропия*, или мера хаоса, беспорядка в системе. Сравнивая выражения I и S, видим, что I можно понимать как информационную *энтропию* (*энтропию* из-за нехватки информации о/в системе).

Л. Больцман дал статистическое определение *энтропии* в 1877 г. и заметил, что *энтропия* характеризует недостающую информацию. Спустя 70 лет, К. Шеннон сформулировал постулаты теории информации, а затем было замечено, что формула Больцмана инвариантна информационной *энтропии*, и была выявлена их системная связь, системность этих фундаментальных понятий.

Важно отметить следующее.

Нулевой *энтропии* соответствует максимальная информация. Основное соотношение между *энтропией* и информацией:

$I+S(\log_2 e)/k=const$

или в дифференциальной форме

$dI/dt= -((\log_2 e)/k)dS/dt.$

При переходе от состояния S_1 с информацией I_1 к состоянию S_2 с информацией I_2 возможны случаи:

1. $S_1 < S_2$ $(I_1 > I_2)$ - уничтожение (уменьшение) старой информации в системе;
2. $S_1 = S_2$ $(I_1 = I_2)$ - сохранение информации в системе;
3. $S_1 > S_2$ $(I_1 < I_2)$ - рождение новой (увеличение) информации в системе.

Главной положительной стороной формулы Шеннона является ее отвлеченность от семантических и качественных, индивидуальных свойств системы. В отличие от формулы Хартли, она учитывает различность, разновероятность состояний - формула имеет статистический характер (учитывает структуру сообщений), делающий эту формулу удобной для практических вычислений. Основной отрицательной стороной формулы Шеннона является то, что она не различает состояния (с одинаковой вероятностью достижения, например), не может оценивать состояния сложных и открытых систем и применима лишь для замкнутых систем, отвлекаясь от смысла информации. Теория Шеннона разработана как теория передачи данных по каналам связи, а *мера Шеннона* - мера количества данных и не отражает семантического смысла.

Увеличение (уменьшение) *меры Шеннона* свидетельствует об уменьшении (увеличении) *энтропии* (организованности) системы. При этом *энтропия* может являться мерой дезорганизации систем от полного хаоса ($S=S^{max}$) и полной информационной неопределенности ($I=I^{min}$) до полного порядка ($S=S^{min}$) и полной информационной определённости ($I=I^{max}$) в системе.

3. ***Термодинамическая мера***. Информационно-термодинамический подход связывает величину *энтропии* системы с недостатком информации о внутренней структуре системы (не восполняемым принципиально, а не просто нерегистрируемым). При этом число состояний определяет, по существу, степень неполноты наших сведений о системе.

Пусть дана термодинамическая система (процесс) S, а H_0, H_1 - термодинамические *энтропии* системы S в начальном (равновесном) и конечном состояниях термодинамического процесса, соответственно. Тогда *термодинамическая мера* информации (негэнтропия) определяется формулой:

$H(H_0,H_1)=H_0 - H_1$.

Эта формула универсальна для любых термодинамических систем. Уменьшение $H(H_0,H_1)$ свидетельствует о приближении термодинамической системы S к состоянию статического равновесия (при данных доступных ей ресурсах), а увеличение - об удалении.

Поставим некоторый вопрос о состоянии термодинамической системы. Пусть до начала процесса можно дать p_1 равновероятных ответов на этот вопрос (ни один из которых не является предпочтительным другому), а после окончания процесса - p_2 ответов. Изменение информации при этом:

$$\Delta I=k \ln(p_1 / p_2)=k (\ln p_1 - \ln p_2).$$

Если $p_1>p_2$ $(\Delta I>0)$ - идет прирост информации, т.е. сведения о системе стали более определенными, а при $p_1<p_2$ $(\Delta I<0)$ - менее определенными. Универсально то, что мы не использовали явно структуру системы (механизм протекания процесса).

Пример. Предположим, что имеется развивающаяся социально-экономическая система с числом состояний 10, которая в результате эволюции развилась до системы с числом состояний 20. Нас интересует вопрос о состоянии некоторого составного элемента системы (например, предприятия). В начале мы знали ответ на вопрос и поэтому $p_1=1$ $(\ln p_1=0)$. Число ответов было пропорционально величине $[\ln 10]$. После развития мы знаем уже микроэкономическое состояние, т.е. изменение информации о состоянии системы равно $\Delta I = -k\ln(20/10) = -k\ln 2$ (нат).

Пример. Предположим, что имеется термодинамическая система - газ в объеме V, который расширяется до объема $2V$.

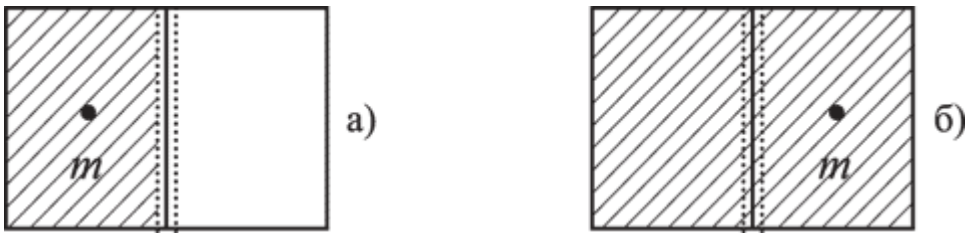

Рис. 6.1. Газ объема V (а) расширяемый до 2V (б)

Нас интересует вопрос о координате молекулы m газа. В начале (а) мы знали ответ на вопрос и поэтому $p_1=1$ $(\ln p_1=0)$. Число ответов было пропорционально $\ln V$. После поднятия заслонки мы уже знаем эту координату (микросостояния), т.е. изменение (убыль) информации о состоянии системы будет равно

$$\Delta I = -k \ln(2V /V) = -k \ln 2 \quad (\text{нат}).$$

Мы получили известное в термодинамике выражение для прироста *энтропии* в расчете на одну молекулу, и оно подтверждает второе начало термодинамики. ***Энтропия*** - мера недостатка информации о микросостоянии статической системы.

Величина ΔI может быть интерпретирована как *количество информации*, необходимой для перехода от одного уровня организации системы к другому (при $\Delta I>0$ - более высокому, а при $\Delta I<0$ - более низкому уровню организации).

Термодинамическая мера (*энтропия*) применима к системам, находящимся в тепловом равновесии. Для систем, далеких от теплового равновесия, например, живых биологических систем, мера-энтропия - менее подходящая.

4. ***Энергоинформационная*** (квантово-механическая) ***мера***. Энергия (ресурс) и информация (структура) - две фундаментальные характеристики систем реального мира, связывающие их вещественные, пространственные, временные характеристики. Если A - именованное множество с носителем так называемого "энергетического происхождения", а B - именованное множество с носителем "информационного происхождения", то можно определить *энергоинформационную меру* $f: A \longrightarrow B$, например, можно принять отношение именования для именованного множества с носителем (множеством имен) A или B. Отношение именования должно отражать механизм взаимосвязей физико-информационных и вещественно-энергетических структур и процессов в системе.

Отметим, что сейчас актуальнее говорить о биоэнергоинформационных мерах, отражающих механизм взаимосвязей биофизико-информационных и вещественно-энергетических структур и процессов в системе.

Пример. Процесс деления клеток сопровождается излучением квантов энергии с частотами приблизительно до $N=1.5\times10^{15}$ гц. Этот спектр можно воспринимать как спектр функционирования словарного запаса клетки как биоинформационной системы. С помощью этого спектра можно закодировать до 1015 различных биохимических реакций, что примерно в 107 раз больше количества реакций реально протекающих в клетке (их количество - примерно 108), т.е. словарный запас клетки избыточен для эффективного распознавания, классификации, регулирования этих реакций в клетке. *Количество информации* на 1 квант энергии: $I=\log_2 10^{15}\approx 50$ бит. При делении клеток количество энергии, расходуемой на передачу 50 бит информации равно энергии кванта (h - постоянная Планка, n - частота излучения):

$$E=h\nu=6{,}62\times10^{-27}\ (\text{эрг/сек})\ \times\ 0{,}5\times10^{15}\ (\text{сек}^{-1})\ =3{,}3\times10^{-12}\ (\text{эрг}).$$

При этом на 1 Вт мощности "передатчика" или на $\mu=10^7$ эрг/сек. может быть передано количество квантов:

$$n=\mu/E=10^7\ (\text{эрг/сек})/(3{,}3\times10^{-12}\ (\text{эрг}))\approx 3{,}3\times10^{18}\ (\text{квант}).$$

Общая скорость передачи информации на 1 Вт затрачиваемой клеткой мощности определяется по числу различных состояний клетки N и числу квантов (излучений) m:

$$V=n\ \log_2 N=3{,}3\times10^{18}\times50\approx1{,}6\times10^{20}\ (\text{бит/сек}).$$

Любая информация актуализируется в некоторой системе. Материальный носитель любой системы - сообщение, сигнал. Любая актуализация сопровождается изменением энергетических свойств (изменением состояния) системы. Наши знания (а, следовательно, и эволюция общества) простираются на столько, на сколько углубляется информация и совершенствуется возможность ее актуализации.

5. Другие меры информации. Многими авторами в последнее время рассматриваются различные количественные меры для измерения смысла информации, например, мера, базирующаяся на понятии цели (А. Харкевич и другие); мера, базирующаяся на понятии тезаурус T=<X,Y,Z>, где X, Y, Z - множества, соответственно, имен, смыслов и значений (прагматики) этих знаний (Ю. Шрейдер и другие); мера сложности восстановления двоичных слов (А. Колмогоров и другие); меры апостериорного знания (Н. Винер и другие); мера успешности принятия решения (Н. Моисеев и другие); меры информационного сходства и разнообразия и другие способы, подходы к рассмотрению мер информации.

Пример. В качестве меры (Колмогорова) восстановления двоичного слова y по заданному отображению f и заданным двоичным словам x из непустого множества X можно взять $H(f,y)=\min|x|$, $x \in X$, $f(x)=y$. Здесь $|x|$ - длина двоичного слова x.

Пример. Если априори известно, что некоторая переменная лежит в интервале (0;1), и апостериори, что она лежит в интервале $(a;b) \subset (0;1)$, тогда в качестве меры (Винера) *количества информации*, извлекаемой из апостериорного знания, можно взять отношение меры $(a;b)$ к мере $(0;1)$.

Пример. В биологических науках широко используются так называемые индексные меры, меры видового разнообразия. Индекс - мера состояния основных биологических, физико-химических и др. компонент системы, позволяющая оценить силу их воздействия на систему, состояние и эволюцию системы. Индексы должны быть уместными, общими, интерпретируемыми, чувствительными, минимально достаточными, качественными, широко применяемыми, рациональными. Например, показателем видового разнообразия в лесу может служить

$$\nu = \sqrt{p_1} + \sqrt{p_2} + \ldots + \sqrt{p_n}$$

где $p_1, p_2, \ldots, p_n$ - частоты видов сообщества, обитающих в лесу, n - число видов.

Вопросы для самоконтроля

1. Что такое мера информации? Каковы общие требования к мерам информации?

2. В чем смысл *количества информации* по Хартли и Шеннону? Какова связь *количества информации* и *энтропии*, хаоса в системе?
3. Какова *термодинамическая мера* информации? Какова квантово-механическая мера информации? Что они отражают в системе?

Задачи и упражнения

1. Система имеет N равновероятных состояний. *Количество информации* в системе (о ее состоянии) равно 5 бит. Чему равна вероятность одного состояния? Если состояние системы неизвестно, то каково *количество информации* в системе? Если известно, что система находится в состоянии номер 8, то чему равно *количество информации*?
2. Некоторая система может находиться в четырех состояниях с вероятностями: в первом (худшем) - 0,1, во втором и третьем (среднем) - 0,25, в четвертом (лучшем) - 0,4. Чему равно *количество информации* (неопределённость выбора) в системе?
3. Пусть дана система с p_0=0,4, p_1=0,5 - вероятности достижения цели управления, соответственно, до и после получения информации о состоянии системы. Оцените меру целесообразности управления этой системой (в битах).

Темы для научных исследований и рефератов

1. *Энтропия* и мера беспорядка в системе. Информация и мера порядка в системе.
2. Квантово-механический и термодинамический подходы к измерению информации.
3. Семантические и несемантические меры информации - новые подходы и аспекты.