Понятие информация

- Информация – одно из самых фундаментальных понятий в современной науке, наряду с веществом, энергией, пространством, временем. А фундаментальное, т.е. первичное, понятие невозможно строго определить через вторичные, или производные понятия.

- Под **информацией в быту** понимают любые сведения об окружающем мире и протекающих в нем процессах, воспринимаемые человеком (с помощью органов слуха, зрения, осязания, обоняния, вкуса) или специальными устройствами.

- Под **информацией в технике** понимают любые сообщения, которые зафиксированы в виде знаков и могут передаваться в виде сигналов.

Под **информацией в теории управления (менеджменте)** понимают сообщения, уменьшающие существующую до этого неопределенность в той предметной области, к которой они относятся, и использующиеся для совершения активного действия, например, управленческого решения.

Под **информацией в теории управления (менеджменте)** понимают сообщения, уменьшающие существующую до этого неопределенность в той предметной области, к которой они относятся, и использующиеся для совершения активного действия, например, управленческого решения.

ключевые атрибуты информации.

- 1. **Достоверность.** информация свободна от ошибок, чьей-либо пристрастности и отражает истинное положение дел. Часто организации применяют независимые источники информации, чтобы анализируя их, уменьшать фактор пристрастности в принимаемом решении или в распространяемой производной информации.
- 2. **Оперативность.** Доставка информации получателям в рамках необходимых временных границ. Например, вчерашняя газета сегодня, запоздавшая котировка акций. Своевременность просто означает, что адресат должен получить информацию, когда ему нужно.
- 3. **Актуальность**, т.е. важность, существенность для настоящего времени. Точная и своевременная информация может в то же время быть неактуальной, более того информация, актуальная для одного получателя, не обязательно актуальна для другого.
- 4. **Полнота.** Информация должна содержать все важные данные, которые ожидают от нее пользователи, и ее должно быть достаточно для понимания и принятия решения.
- 5. **Полезность.** Полезность (ценность) информации определяется по тем задачам, которые можно решить с ее помощью.
- 6. **Понятность** означает, что информация может быть представлена в ясном и понятном для потребителя формате. Потребитель информации – лицо, принимающее решение, должен как можно меньше времени тратить на дополнительные уточнения поступившей информации.

- в узком смысле информацией можно назвать сведения о предметах, фактах, понятиях некоторой предметной области.
- С середины XX века **информация** рассматривается в **широком смысле** как общенаучное понятие, включающее в себя как совокупность сведений об объектах и явлениях окружающей среды, их параметрах, свойствах и состоянии, так и обмен сведениями между людьми, человеком и автоматом, автоматом и автоматом, обмен сигналами между живой и неживой природой, в животном и растительном мире, а также генетическую информацию.

информацию можно подразделить на:

1) структурную (или связанную) присущую объектам неживой и живой природы естественного или искусственного происхождения. Эти объекты (орудия труда, предметы быта, произведения искусства, научные теории и т.п.) возникают путем опредмечивания циркулирующей информации, то есть благодаря и в результате целенаправленных управленческих процессов;

2) оперативную (или рабочую), циркулирующую между объектами материального мира и используемую в процессах управления в живой природе, в человеческом обществе

Данные, знания

- Сведения, полученные путем измерения, наблюдения, логических или арифметических операций, и представленные в форме, пригодной для постоянного хранения, передачи и обработки получили название **данные**.
- Совокупность полезной информации, правил и процедур ее обработки, необходимая для получения новой информации о какой-либо предметной области называют **знанием**.

Свойства знаний

- 1. Внутренняя интерпретируемость знаний (понятность знания его носителю).
- 2. Структурированность знаний. Информационные единицы должны обладать гибкой структурой. Принцип «матрешки» – рекурсивная вложимость знаний. Возможность произвольного установления и перенастройки отношений (включения) между информационными единицами.
- 3. Связность. Отношения между элементами: структурные, функциональные, казуальные и семантические. Структурные задают иерархию, функциональные задают процедурную информацию, позволяющие находить одни элементы через другие, каузальные задают причинно-следственные связи, семантические охватывают все остальные виды отношений.
- 4. Ассоциативность знаний – наличие семантической метрики в сфере знаний. Отношение релевантности на множестве информационных единиц характеризует ситуационную близость элементов (силу ассоциативной связи). Позволяет находить знания, близкие к уже
- найденным.
- 5. Активность знаний – наличие у знаний побуждающей и направляющей функции, что фактически превращает знания в квазипотребности. Актуализации тех или иных действий способствуют имеющиеся в системе знания.

- Введение информации в научно-технический и хозяйственный оборот привело к необходимости ее количественной оценки, т.е. к введению меры сравнения. В простейшей комбинаторной форме эта мера была предложена Р. Хартли в 1928 году.

- **Пример 1**. Как определить, какая из двух монет фальшивая, если на вид они одинаковы, но известно, что фальшивая легче. Нет ничего проще, скажете вы. Проводим одно взвешивание на чашечных весах, и все становится ясно. Таким образом, до взвешивания у вас *была неопределенность* по поводу того, какая из монет фальшивая, а после взвешивания вы *сняли эту неопределенность*, получив *информацию*. Иными словами, вы получили сообщение в элементарном альтернативном выборе между двумя событиями («фальшивая – не фальшивая», «да - нет», «истина – ложь», «0–1»).

Понятие бита

- бит – это и двоичный знак, и единица измерения количества информации, определяемая как ***количество информации в выборе с двумя взаимоисключающими равновероятными исходами***.

- **Пример 2.** А если монет 8? Тогда делим их на две равные части и взвешиваем их. Ту часть, которая легче, снова делим на две части и снова взвешиваем и т.д. За три взвешивания мы определим фальшивую монету.
- За три выбора мы уменьшили существующую неопределенность в 2, 4, 8 раз, получив таким образом 3 бита информации.

Пример 3. Перейдем от монет к картам. Пусть в колоде из 32 карт необходимо угадать определенную карту, например, туза пик. Для этого необходимо и достаточно получить ответы «да» и «нет» на **пять** вопросов. Вопросы, ответы на которые позволяют выбрать одну из альтернатив, называют двоичными, или бинарными. Ответами на эти вопросы мы уменьшаем неопределенность в 2, 4, 8, 16, 32 раз. В конце неопределенности не остается. Количество полученной информации равно 5 бит

Вопрос	Ответ	Бинарный ответ
1.Карта красной масти?	Нет	0
2.Трефы?	Нет	0
3.Одна из четырех старших?	Да	1
4.Одна из двух старших?	Да	1
5.Король?	Нет	0
Значит, задуманная карта была туз пик.		

- В этих примерах процесс получения информации рассматривается как выбор одного сообщения из конечного наперёд заданного множества из *n* **равновероятных** сообщений. Легко подметить следующую закономерность: количество информации *I*, содержащееся в выбранном сообщении, определяется как двоичный логарифм *n*:

$$I = \mathrm{ld}(n)$$

- Справедливо утверждение Хартли: если во множестве $X=\{x1,x2,\ldots,xn\}$ выделить произвольный элемент $xi \in X$, то, чтобы его найти, необходимо получить не менее ld(n) единиц информации.
- Недостаток формулы Хартли заключается в том, что она не учитывает **неравновероятность** различных рассматриваемых состояний.

Вероятности отдельных букв в русском языке (с учетом пробела)

Буква	—	О	Е,Ё	А	И	Т	Н	С
P_i	0,175	0,090	0,072	0,062	0,062	0,053	0,053	0,045
Буква	Р	В	Л	К	М	Д	П	У
P_i	0,040	0,038	0,035	0,028	0,026	0,025	0,023	0,021
Буква	Я	Ы	З	Ь,Ъ	Б	Г	Ч	Й
P_i	0,018	0,016	0,016	0,014	0.014	0,013	0,012	0,010
Буква	Х	Ж	Ю	Ш	Ц	Щ	Э	Ф
P_i	0,009	0,007	0,006	0,006	0,004	0,003	0,003	0,002

Частоты букв (в процентах) ряда европейских языков

Буква алфавита	Французский язык	Немецкий язык	Английский язык	Итальянский язык
A	7,68	5,52	7,96	11,12
B	0,80	1,56	1,60	1,07
C	3,32	2,94	2,84	4,11
D	3,60	4,91	4,01	3,54
E	17,76	19,18	12,86	11,63
F	1,06	1,96	2,62	1,15
G	1,10	3,60	1,99	1,73
H	0,64	5,02	5,39	0,83
I	7,23	8,21	7,77	12,04
J	0,19	0,16	0,16	-
K	-	1,33	0,41	-
L	5,89	3,48	3,51	5,95
M	2,72	1,69	2,43	2,65
N	7,61	10,20	7,51	7,68
O	5,34	2,14	6,62	8,92
P	3,24	0,54	1,81	2,66

- Для неравновероятных процессов американский учёный Клод Шеннон предложил (1948 г.) другую формулу определения количества информации, которая учитывает возможную неодинаковую вероятность сообщений во множестве сообщений.
- *Вероятность* – это численная мера достоверности случайного события, которая при большом числе испытаний близка к отношению числа случаев *m*, когда событие осуществилось (положительных исходов), к общему числу случаев *n*:

$$P_i = \lim_{n \to \infty}(m / n)$$

Например, если много раз подбрасывать монетку, то она упадёт орлом вверх примерно в половине случаев. Это значит, что вероятность выпадения орла равна 0,5, или 50%.

Вероятность любого события – это число, принадлежащее отрезку [0;1]. Событие с вероятностью 0 называют невозможным, а с вероятностью 1 – достоверным.

Свойство вероятностей:

$$\sum_{i=1}^{n} P_i = 1$$

- Можно представить что для того, чтобы получить какой то символ от источника сообщения нужно перебрать по крайней мере (с вероятностью близкой к 1) n символов, где n=1/p. p –вероятность появления символа. Чтобы получить из этих символов необходимый нам, нужно сделать двоичный логарифм переборов ld(n).

$$K_i = \mathrm{ld}(1/P_i) \quad \text{бит}$$

мера Шеннона количества информации (формула Шеннона).

•Если найти среднее значение количества таких переборов для всех символов получим формулу Шеннона:

$$H = \sum_{i=1}^{n} P_i I_i = \sum_{i=1}^{n} P_i \,\mathrm{ld}\left(\frac{1}{P_i}\right) = -\sum_{i=1}^{n} P_i \,\mathrm{ld}(P_i)$$

Эта величина получила название ***информационная энтропия***, или энтропия источника сообщений.

Энтропия характеризует информационную мощность данного множества (ансамбля) сообщений и является мерой неопределенности, которая имеется в этом множестве.

$$H = \sum_{i=1}^{n} P_i I_i = \sum_{i=1}^{n} P_i \,\mathrm{ld}\left(\frac{1}{P_i}\right) = -\sum_{i=1}^{n} P_i \,\mathrm{ld}(P_i)$$

- Из формулы непосредственно вытекают свойства энтропии:
- энтропия заранее известного сообщения равна 0;
- во всех других случаях $H > 0$.

Чем больше энтропия системы, тем больше степень ее неопределенности. Поступающее сообщение полностью или частично снимает эту неопределенность. Поэтому *количество информации можно измерять тем, насколько понизилась энтропия системы после поступления сообщения*:

Уменьшая энтропию, мы получаем информацию – в этом и заключается смысл научного познания!

Кодирование источника сообщений

- Как уже отмечалось, результат одного отдельного альтернативного выбора может быть представлен как 0 или 1. Тогда выбору всякого сообщения (события, символа т.п.) в массиве сообщений соответствует некоторая последовательность двоичных знаков 0 или 1, то есть *двоичное слово*. Это двоичное слово называют *кодировкой*, а множество кодировок источника сообщений – *кодом источника сообщений*.

- Если количество символов представляет собой степень двойки ($n = 2^N$) и все знаки равновероятны $P_i = (1/2)^N$, то все двоичные слова имеют длину $L=N=\mathrm{ld}(n)$. Такие коды называют ***равномерными кодами***.
- Более оптимальным с точки зрения объема передаваемой информации является ***неравномерное кодирование***, когда разным сообщениям в массиве сообщений назначают кодировку разной длины. Причем, часто происходящим событиям желательно назначать кодировку меньшей длины и наоборот, т.е. учитывать их вероятность.

Кодирование словами постоянной длины

Буква	*a*	*b*	*c*	*d*	*e*	*f*	*g*
Кодирование	000	001	010	011	100	101	110

ld(7)≈2,807 и *L*=3.

. Проведем кодирование, *разбивая исходное множество знаков на равновероятные подмножества*, то есть так, чтобы при каждом разбиении суммы вероятностей для знаков одного подмножества и для другого подмножества были одинаковы. Для этого сначала расположим знаки в порядке уменьшения их вероятностей

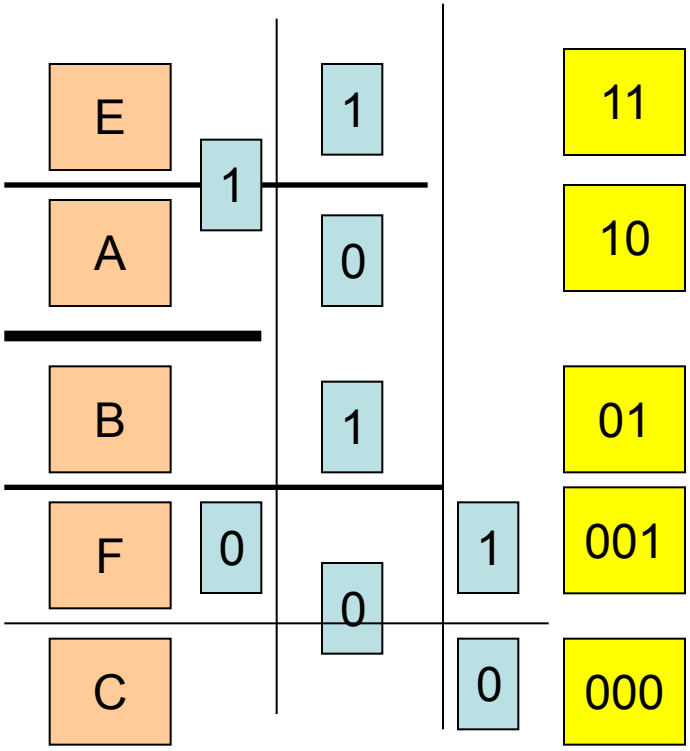
Символ	Вероятность, P_i	Кодировка	Длина, L_i	Вероятность×Длина, $P_i \times L_i$
a	0,25	00	2	0,5
e	0,25	01	2	0,5
f	0,125	100	3	0,375
c	0,125	101	3	0,375
b	0,125	110	3	0,375
d	0,0625	1110	4	0,25
g	0,0625	1111	4	0,25

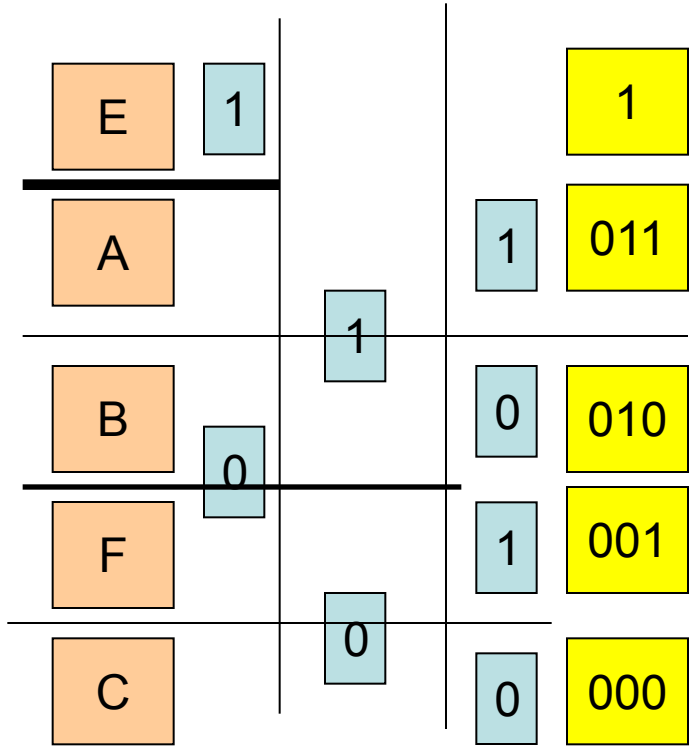
В общем случае ***алгоритм построения оптимального кода Шеннона-Фано*** выглядит следующим образом:

1. сообщения, входящие в ансамбль, располагаются в столбец по мере убывания вероятностей;
2. выбирается основание кода K (в нашем случае K=2);
3. все сообщения ансамбля разбиваются на K групп с суммарными вероятностями внутри каждой группы как можно близкими друг к другу.
4. всем сообщениям первой группы в качестве первого символа присваивается 0, сообщениям второй группы – символ 1, а сообщениям K-й группы – символ (K–1); тем самым обеспечивается равная вероятность появления всех символов 0, 1,…, K на первой позиции в кодовых словах;
5. каждая из групп делится на K подгрупп с примерно равной суммарной вероятностью в каждой подгруппе. Всем сообщениям первых подгрупп в качестве второго символа присваивается 0, всем сообщениям вторых подгрупп – 1, а сообщениям K-х подгрупп – символ (K–1).
6. процесс продолжается до тех пор, пока в каждой подгруппе не окажется по одному сообщению.

Символ	Вероятность, P_i
A	0,2
E	0,4
F	0,125
C	0,125
B	0,15

Символ	Вероятность, P_i
E	0,4
A	0,2
B	0,15
F	0,125
C	0,125





При неравномерном кодировании вводят среднюю длину кодировки, которая определяется по формуле

$$L = \sum_{i=1}^{n} P_i L_i$$

В общем же случае связь между средней длиной кодового слова *L* и энтропией *H* источника сообщений дает следующая ***теорема кодирования*** Шеннона:

имеет место неравенство $L \geq H$, причем $L = H$ тогда, когда набор знаков можно разбить на точно равновероятные подмножества;

всякий источник сообщений можно закодировать так, что разность $L - H$ будет как угодно мала.

Разность *L–H* называют *избыточностью кода* (мера бесполезно совершаемых альтернативных выборов).

следует *не просто кодировать каждый знак в отдельности*, а рассматривать вместо этого двоичные кодирования для *nk* групп по *k* знаков.

Тогда средняя длина кода *i*-го знака *xi* вычисляется так:

L = (средняя длина всех кодовых групп, содержащих *xi*)/*k*.

Символ	Вероятность	Кодировка	Длина	В×Д
A	0,7	0	1	0,7
B	0,2	10	2	0,4
C	0,1	11	2	0,2

Средняя длина слова: $L = 0,7+0,4+0,2=1,3$.

Среднее количество информации, содержащееся в знаке (энтропия):

$H = 0,7\times\mathrm{ld}(1/0,7)+0,2\times\mathrm{ld}(1/0,2)+0,1\times\mathrm{ld}(1/0,1) = 0,7\times0,515+0,2\times2,322+$
$+0,1\times3,322 = =1,1571$.

Избыточность $L - H = 1,3 - 1,1571 = 0,1429$.

Кодирование пар

Пары	Вероятность	Кодировка	Длина	В×Д
AA	0,49	0	1	0,49
AB	0,14	100	3	0,42
BA	0,14	101	3	0,42
AC	0,07	1100	4	0,28
CA	0,07	1101	4	0,28
BB	0,04	1110	4	0,16
BC	0,02	11110	5	0,10
CC	0,02	111110	6	0,12

Средняя длина кода одного знака равна 2,33/2=1,165 – уже ближе к энтропии.
Избыточность равна $L - H = 1{,}165 - 1{,}1571 \approx 0{,}008$.