# Измерение информации

Поскольку существуют различные определения информации, есть и различные подходы к ее измерению.

## Технический (объемный) подход

Наиболее простой подход, измеряющий количество информации как число знаков (сигналов) в сообщении, без учета ее смысла.

*Какой это подход к определению информации?*

Основная единица измерения – **бит**. Большое распространение получил **байт** (8 бит) – минимальная единица адресации памяти в компьютере.

**Двоичные приставки**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Степени 10** | | | **Степени 2** | | | |
| СИ/ГОСТ | | |  | МЭК | | ГОСТ |
| 103Б | кило- | кБ | 210Б = 1024 Б | киби- | КиБ | Кбайт |
| 106Б | мега- | МБ | 220Б ≈ 1,05∙106Б | меби- | МиБ | Мбайт |
| 109Б | гига- | ГБ | 230Б ≈ 1,07∙109Б | гиби- | ГиБ | Гбайт |
| 1012Б | тера- | ТБ | 240Б ≈ 1,10∙1012Б | теби- | ТиБ | Тбайт |
| 1015Б | пета- | ПТ | 250Б ≈ 1,13∙1015Б | пеби- | ПиБ | Пбайт |
| 1018Б | экса- (экза-) | ЭБ | 260Б ≈ 1,15∙1018Б | эксби- | ЭиБ | Эбайт |
| 1021Б | зетта- | ЗБ | 270Б ≈ 1,18∙1021Б | зеби- | ЗиБ | Збайт |
| 1024Б | йотта- | ЙБ | 280Б ≈ 1,21∙1024Б | йоби- | ЙиБ | Йбайт |

В общем случае могут использоваться любые удобные единицы измерения – страницы, книги, DVD-диски, экраны.

Например, в полиграфии используется **печатный лист** (п.л.) стандартного размера. Объем издания зависит от размера шрифта, полей, иллюстраций и т.п.

Рукописи измеряются в **авторских листах** (а.л.) – 40 000 знаков, включая пробелы. Объем не зависит от оформления, не учитывает рисунки.

Преобразование объема к двоичным единицам зависит от **кодировки** – сколько бит необходимо, чтобы закодировать алфавит.

Пример 1. Таблица ASCII содержит 256 символов. Какой объем текста в битах можно вывести на экран терминала, если он содержит 25 строк по 80 символов?

Пример 2. В Unicode каждый символ кодируется двумя байтами. Сколько символов содержит таблица Unicode?

Пример 3. Каждый пиксель изображения кодируется тремя байтами. Сколько изображений размером 4000x3000 пикс. можно записать на стандартный DVD?

В общем случае, если все символы алфавита кодируются одинаковым числом бит



i – информационный объем символа;

N – количество символов в алфавите (мощность алфавита);

 – округление вверх.

Пример 4. Сколько бит нужно, чтобы закодировать один символ алфавита, включающего:

* строчные и заглавные буквы русского алфавита;
* арабские цифры;
* знаки препинания . , ! ? ... : ; «» - –
* скобки () [] {};
* пробел и знак «конец абзаца».

Сколько полных байт займет текст из 20 символов этого алфавита?

Пример 5. Вождь племени Мумба отправил племени Юмба глиняную табличку с посланием из 80 букв. В ответ он получил свиток с 23 иероглифами. Чье сообщение больше, если в алфавите племени Мумба 29 букв, а иероглифов у племени Юмба около 1500?

## Вероятностный (энтропийный) подход

Используется в теории информации.

Получение информации предполагает увеличение объема знания и уменьшение незнания, неопределенности – **энтропии**.

За **единицу (бит) количества информации** принимается выбор одного из двух равновероятных вариантов («да» или «нет», «1» или «0», «ложь» или «истина»).

Пример 1. Монета может упасть «орлом» или «решкой» с равной вероятностью. Узнав результат одного броска, вы получите 1 бит информации.

Пример 2. Вы определенно помните, что Иванов живет на улице Лютиков, но не помните номер дома. Немного подумав, вы вспомнили, что номер его дома четный. Сколько информации вы получили?

**Формула Хартли:**



*I* – информационный объем сообщения;

*N* – количество равновероятных вариантов выбора.

Если *N* = 2 вариантам, то *I* = 1 бит.

Формула совпадает с формулой для объема 1 символа в объемном подходе. Т.е. можно сказать, что мы выбираем 1 из *N* возможных символов алфавита с равной вероятностью.

Иначе можно записать:

,

*p* – вероятность появления одного из вариантов.

Пример 3. В колоде 32 карты. Сколько информации вы получите, узнав, что вытянули карту:

а) черного цвета;

б) масти пик;

в) десятку;

г) даму бубен?

Для неравных вероятностей выбора:

,

*hi* – информативность *i*-того варианта (**частная энтропия**);

*pi* – вероятность появления *i*-того варианта.

Пример 4. Покупатель может приобрести или не приобрести определенный товар. Но вероятность этих событий неодинакова. Поэтому узнав, купил он товар, или не купил, вы получите разное количество информации, не равное 1 биту.

Пусть вероятность покупки составляет 80%, а «не покупки» – 20%. Какое сообщение информативнее:

а) покупатель купил товар;

б) покупатель не купил товар?

Т.е., чем меньше вероятность появления варианта выбора, тем больше его энтропия, и в то же время информативность.

В общем случае вероятность появления символов алфавита неодинакова. Например, в русском языке чаще всего встречаются буквы «а» и «о». Заглавные буквы встречаются реже строчных.

**Формула Шеннона** – энтропия информационного сообщения:

,

*H* – средняя информативность одного символа в сообщении.

Свойства энтропии:

1. H ≥ 0
2. 

Если вариант выбора всего один, то :

.

Т.е. неопределенность отсутствует.

Пример 5. Оцените для себя вероятность получения каждой оценки на следующей самостоятельной. Сколько бит информации несет каждый вариант? Какова средняя информативность оценки за самостоятельную?

Пример 6.

Известно, что среди покупателей магазина 63% приобретают товар марки А, а 21% – товар марки Б. Чему равна энтропия каждого варианта покупки? Общий объем информации о потребительских предпочтениях 100 покупателей?

PA = 0,63

PB **= 0,21**

**P0 =0,16**

**HA =**

На знании энтропии символов основано *энтропийное кодирование*, позволяющее сжимать физический объем данных за счет переменной длины кода для каждого символа. Кратко говоря, при этом используются короткая запись для часто встречающихся символов и длинная – для редко встречающихся.

## Семантический (смысловой) подход

Наиболее трудно формализуемый. Связан с субъективной, эмоциональной оценкой полученной информации. Оценивается новизна, интересность информации. Оценку дает *получатель,* поэтому объемзависит не только от самой информации, но и от качеств получателя. Оцениваться может и синтаксис, и семантика.

Пример. На вопрос «Ты меня любишь?» можно получить два варианта ответа «да» или «нет». Но смысл ответа не оценивается 1 битом информации.

Один из известных научных вариантов – **тезаурусная мера** (Ю.И. Шнейдер). Тезаурус (словарь) – совокупность сведений, которыми располагает человек или система. Тогда количество смысловой информации в сообщении можно оценить степенью изменения тезауруса после получения этого сообщения.

## Прагматический подход

Прагматическая мера связана с определением количества информации, необходимым для достижения поставленной цели и определяет степень ценности, полезности информации. Ценность информации обычно измерять в тех же единицах, что и целевая функция (руб. в экономике, м/с – скорость космического аппарата, оценка для студента, число голосов на выборах и т.д.).

Например, в экономической системе прагматическую ценность информационной технологии можно определить приростом экономического эффекта от внедрения этой технологии.

## Информационный взрыв

По оценкам экспертов, объем сгенерированных данных составил в 2011г. 1,8ЗБ, в 2012г. 2,8ЗБ, к 2020г., может достигнуть 40ЗБ. При этом общий объем информации будет удваиваться каждые 18 месяцев. По большей части этот поток состоит из неструктурированных данных (лишь 5 % составляют различные базы данных – тем или иным образом структурированная информация).

**Информационный взрыв** — постоянное увеличение скорости и объёмов публикаций (объёма информации) в масштабах планеты.

Первые жалобы на слишком большой объем информации можно найти еще до нашей эры, в Древней Греции и Китае.

С начала нашей эры первое удвоение накопленных знаний произошло в 1750 году. Второе уже в 1900г. (через 150 лет), третье в 1950г. (через 50 лет). Сейчас за один день человечество сейчас производит больше информации, чем в 1986 году накопило за всю свою историю.

**Информационные революции:**

1. Письменность
2. Книгопечатание
3. Электричество (телеграф, телефон)
4. Микропроцессоры и компьютерные сети

Парадокс:

Использование поисковых машин позволяет пользователям быстро находить информацию. Тем не менее, информация, опубликованная онлайн, не всегда надежна из-за отсутствия полномочий утверждения и обязательной точности проверки перед публикацией. В результате люди пользуются перекрестной проверкой читаемых материалов перед принятием решения, и это занимает больше времени.

Избыток информации приводит к **информационной перегрузке** и замедляет принятие решения.

Основные причины информационной перегрузки:

* стремительный рост объема новой информации;
* простота дублирования и передачи данных через интернет;
* увеличение доступных каналов входящей информации (например, телефон, электронная почта, мгновенный обмен сообщениями, rss);
* большое количество исторических сведений;
* противоречия и неточности в имеющейся информации;
* низкую долю полезной информации в общем потоке;
* отсутствие метода сравнения и обработки различных видов информации;
* куски информации не связаны или не имеют общей структуры для выявления их отношений.