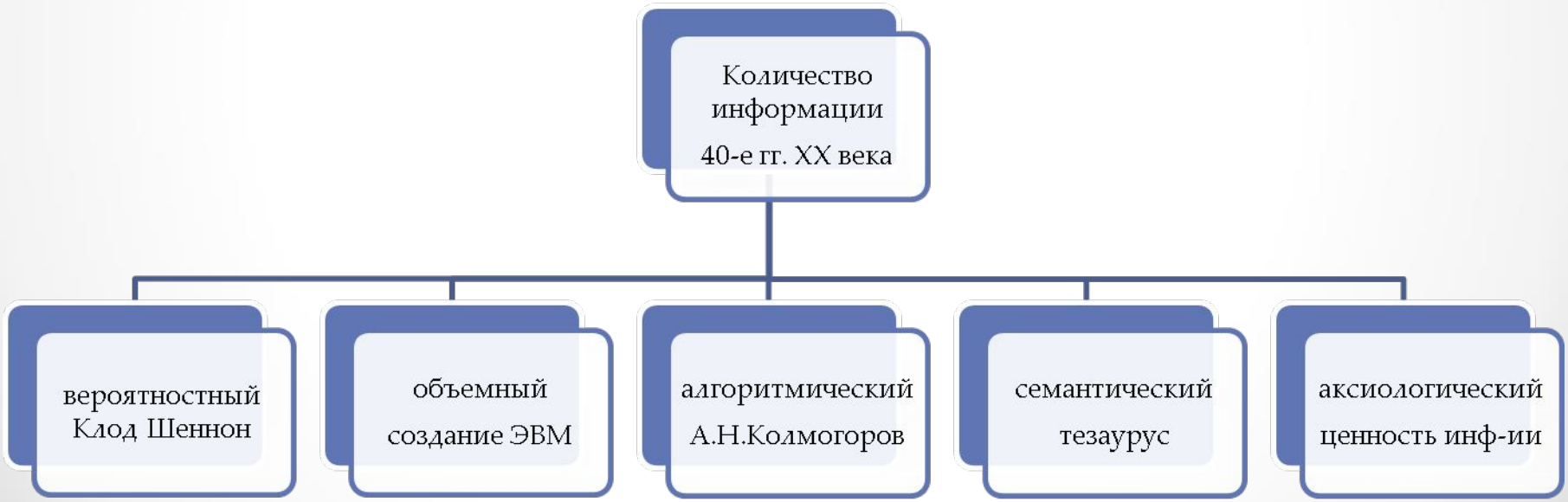
§ 4. Количество информации

Подходы к определению понятия «количество информации»



- **Алгоритмический**: любому сообщению можно приписать количественную характеристику, отражающую сложность (размер) программы, которая позволяет его произвести.
- **Семантический**: тезаурус – совокупность сведений, которыми располагает пользователь или система. Количество семантической информации зависит от соотношения между смысловым содержанием и тезаурусом.
- **Аксиологический**: исходит из ценности, практической значимости информации, качественных характеристик, значимых в социальной среде.

Объемный подход

Создатели компьютеров отдали пред-почтение двоичной системе счисления, т.к. в техническом устройстве наиболее просто реализовать два противоположных физических состояния.

Наименьшая единица информации – **бит** (BInary digiTs).

Бит – это ответ на вопрос, требующий односложного разрешения – да или нет.

Объемный подход

1 байт = 8 бит

1 килобайт = 1024 байта = 2^{10} байт

2 Кб – одна страница неформатированного машинного текста

1 мегабайт = 1024 килобайта = 2^{20} байт

1 гигабайт = 1024 мегабайта = 2^{30} байт

1 Терабайт = 1024 гигабайта = 2^{40} байт

1 Тб – 15 фильмов среднего качества

Вероятностный (энтропийный) подход

- принят в теории информации и кодирования
- получатель сообщения имеет определенное представление о возможных наступлениях некоторых событий (выражаются **вероятностями**, с которыми он ожидает то или иное событие). Получаемая информация уменьшает число возможных вариантов выбора (т.е. неопределенность), а полная информация не оставляет вариантов вообще.
- **Энтропия** – общая мера неопределенностей. Количество информации в сообщении = насколько уменьшилась эта мера после получения сообщения

Вероятность

Идет ли сейчас снег?

Вероятность – это число в интервале от 0 до 1.

p=1 – событие обязательно произойдет

p=0 – событие никогда не произойдет

Бросаем монетку и смотрим: «орел» или «решка». Если повторять этот опыт много раз, то количество «орлов» и «решек» примерно равно. Вероятность каждого из двух событий равна 0,5.

Классический игральный кубик вероятность 1/6

Вероятностный (энтропийный) подход

Американский инженер Ральф Хартли в 1928г. предложил формулу

$$I = \log_2 N$$

N – количество вариантов

I – количество информации (в битах)

1 бит – количество полученной информации при выборе из двух возможных вариантов

Для значений N, не равных целой степени числа 2, количество информации – дробное число.

Пример

Американский инженер Ральф Хартли в 1928г. предложил формулу

$$I = \log_2 N$$

N – количество вариантов

I – количество информации (в битах)

1 бит – количество полученной информации при выборе из двух возможных вариантов

Для значений N, не равных целой степени числа 2, количество информации – дробное число.

Теоретическое количество информации в сообщении

Американский инженер Ральф Хартли в 1928г. предложил формулу

$$I = \log_2 N$$

N – количество вариантов

I – количество информации (в битах)

1 бит – количество полученной информации при выборе из двух возможных вариантов

Для значений N, не равных целой степени числа 2, количество информации – дробное число.

Алфавитный подход

Американский инженер Ральф Хартли в 1928г. предложил формулу

$$I = \log_2 N$$

N – количество вариантов

I – количество информации (в битах)

1 бит – количество полученной информации при выборе из двух возможных вариантов

Для значений N, не равных целой степени числа 2, количество информации – дробное число.

Алфавитный подход

- на практике используют первое целое число, которое больше теоретически рассчитанного;
- все события (символы алфавита) одинаково ожидаемы;
- смысл сообщения не учитывается;
- в реальности это предположение не всегда верно (например, в тексте на русском языке);
- Такой подход (**важен только объем**) очень удобен для устройств, передающих информацию по сети;
- Чаще всего применяют для вычисления информационного объема текста.

Частотность букв русского языка

Пробел	0,175	**Я**	0,018
О	0,090	**Ы**	0,016
Е, Ё	0,072	**З**	0,016
А	0,062	**Ь, Ъ**	0,014
И	0,062	**Б**	0,014
Т	0,053	**Г**	0,013
Н	0,053	**Ч**	0,012
С	0,045	**Й**	0,010
Р	0,040	**Х**	0,009
В	0,038	**Ж**	0,007
Л	0,035	**Ю**	0,006
К	0,028	**Ш**	0,005
М	0,026	**Ц**	0,004
Д	0,025	**Щ**	0,003
П	0,023	**Э**	0,003
У	0,021	**Ф**	0,002

Понятие вероятности

американский инженер Ральф Хартли в 1928г. предложил формулу

$$I = \log_2 N$$

N – количество вариантов

I – количество информации (в битах)

1 бит – количество полученной информации при выборе из двух возможных вариантов

Для значений N, не равных целой степени числа 2, количество информации – дробное число.

Классический игральный кубик имеет 6 граней.

Вероятность выпадения каждой грани равна 1/6.

Вероятность выпадения четного числа равна 0,5.

Вероятность выпадения числа, меньшего 3, равна 1/3.

Вероятностный (энтропийный) подход

- Американский инженер Ральф Хартли в 1928г. предложил формулу

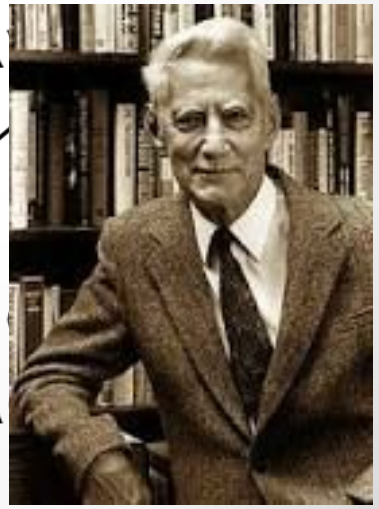
$$I = \log_2 N$$

N – количество вариантов

I – количество информации (в битах)

1 бит – количество пол информации при выборе и возможных вариантов

Для значений N, не равных целой числа 2, количество информ дробное число.



Понятие энтропии

Энтропия –

- в естественных науках - мера беспорядка системы, состоящей из многих элементов
- в теории информации — мера неопределённости какого-либо опыта (испытания), который может иметь разные исходы, а значит, и количество информации

Явление, обратное энтропии, именуется **негэнтропией.**

Понятие энтропии

- Американский инженер Ральф Хартли в 1928г. предложил формулу

$$I = \log_2 N$$

N – количество вариантов

I – количество информации (в битах)

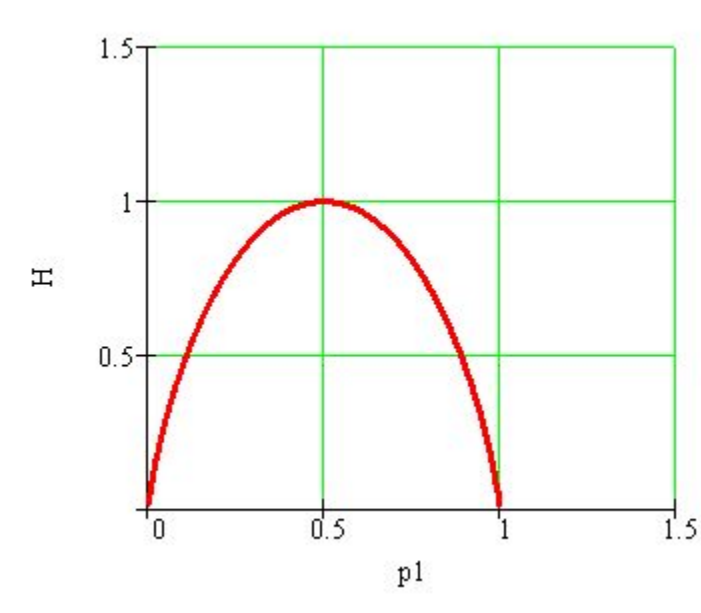
1 бит – количество полученной информации при выборе из двух возможных вариантов

Для значений N, не равных целой степени числа 2, количество информации – дробное число.

Понятие энтропии

Два события: «Снег идет» и «Снега нет» – составляют полную систему.

Сумма вероятностей всех событий, составляющих полную систему, **равна 1**.



Неопределенность наибольшая для случая, когда все события равновероятны.

При равновероятных событиях неопределенность совпадает с количеством информации, вычисленной по формуле Хартли.

Пример задачи

Определите количество информации в сообщении **с учетом** и **без учета вероятности** появления символов в сообщении, **определите энтропию и избыточность алфавита** в сообщении. Точность вычисления – три знака после запятой. Сообщение – КАРАВАН.

Решение:

1. Найдём количество информации без учёта вероятности по формуле:

$$I_{б.у.} = \log_2 N = \log_2 7 = 2{,}807 \text{ бит},$$

где N – общее число символов в сообщении.

Пример задачи

2. Найдём количество информации с учётом вероятности. Найдем вероятность появления каждой буквы:

$$p_к = \frac{1}{7}, \quad p_а = \frac{3}{7}, \quad p_р = \frac{1}{7}, \quad p_в = \frac{1}{7}, \quad p_н = \frac{1}{7}$$

Определим количество информации для каждой буквы в сообщении по формуле:

$$i_{буква} = \log_2 \frac{1}{p_{буква}}$$

$$i_к = i_р = i_в = i_н = \log_2 \frac{1}{1/7} = \log_2 7 = 2{,}807$$

$$i_а = \log_2 \frac{7}{3} = 1{,}222$$

Количество информации всего сообщения:

$$I_{с.у} = \sum_{k=1}^{N} i_k = 4 * 2.807 + 3 * 1.222 = 14.894$$

Пример задачи

3. Определим энтропию сообщения:

$$H(\alpha) = -\sum_{i=1}^{n} p_i * \log_2 p_i = -\left(\frac{1}{7} * \log_2 \frac{1}{7} * 4 + \frac{3}{7} * \log_2 \frac{3}{7}\right) = 2{,}132$$

4. Определим избыточность символов:

$$D = \frac{H_{\max}(\alpha) - H(\alpha)}{H_{\max}(\alpha)} = \frac{I_{б.у} - H(\alpha)}{I_{б.у}} = \frac{2.807 - 2.132}{2.807} = 0.24$$