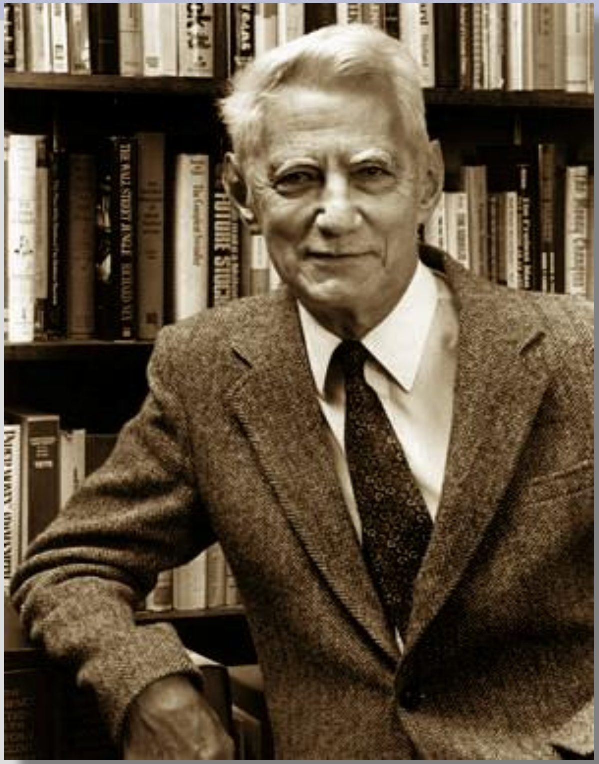
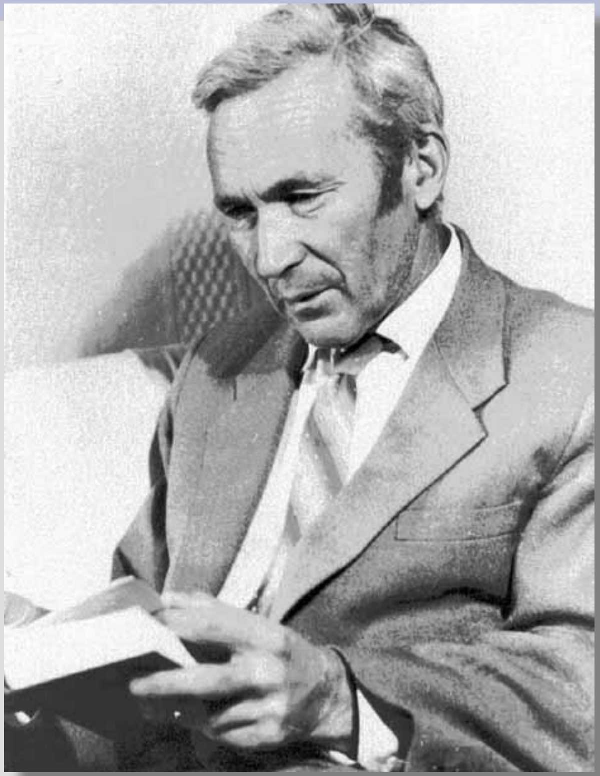
ОСНОВЫ ТЕОРИИ ИНФОРМАЦИИ

Понятие информации. Количество информации.
Урок 1.

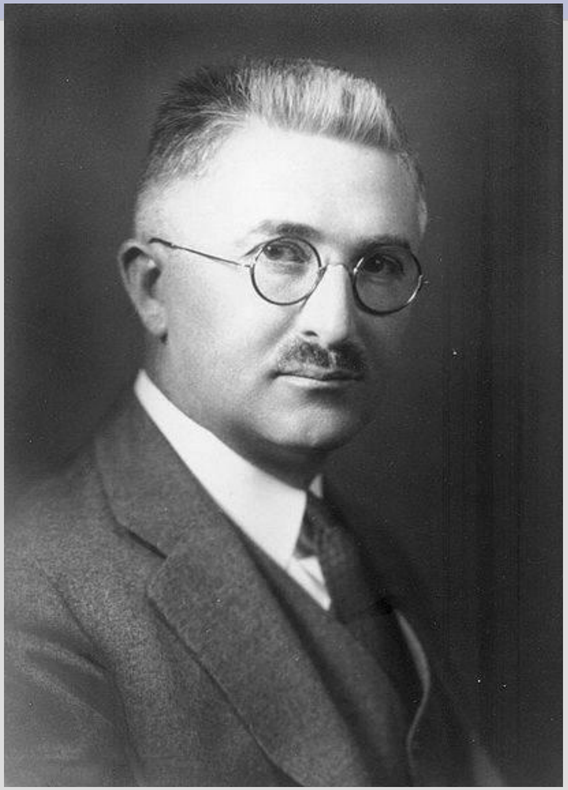
Понятие информации



Клод Э́лвуд Ше́ннон (США)



Андре́й Никола́евич Колмого́ров (Россия)



Ральф Винтон Лайон Хартли (США)

- **Теория информации -** вторая половина XX века

Понятие информации

- **Информация** — это *сведения*, обладающие такими характеристиками, как понятность, достоверность, новизна и актуальность

(*с точки зрения человека*)

- **Информация** — это снятая *неопределенность*

(по К. Шеннону)

- **Количество информации** - минимально возможное количество двоичных знаков, необходимых для кодирования последовательности к содержанию представленного сообщения

(по А.Н. Колмогорову)

- **Информационный объем** сообщения — количество двоичных символов, которое используется для кодирования этого сообщения.

Единицы измерения информации

За единицу измерения информации принят
1 бит (от англ. **Bi**nary digi**t**)

1 бит — это количество информации, которое можно передать в сообщении, состоящем из одного двоичного знака (0 или 1) — *алфавитный подход*

1 бит — это количество информации, уменьшающее неопределенность знаний в два раза — *содержательный подход*

1Кб (килобайт) = 2^{10} байт = 1024 байт
1Мб (мегабайт) = 2^{10} Кб = 1024 Кб
1Гб (гигабайт) = 2^{10} Мб = 1024 Мб
1Тб (терабайт) = 2^{10} Гб =1024 Гб

ВОПРОСЫ И ЗАДАНИЯ

1. Расскажите, как вы понимаете термин «информация». Что общего и каковы различия между бытовым понятием этого термина и его научными трактовками?
2. При игре в кости используется два игральных кубика, грани которых помечены числами от 1 до 6. В чем заключается неопределенность знаний о бросании одного кубика? Двух кубиков одновременно?
3. Сколько гигабайт содержится в 2^{18} килобайтах?
4. Сколько мегабайт содержится в 2^{20} килобитах?

Формула Хартли определения информации.
Урок 2.

Задача 1. *Предположим, что в классе находятся 32 ученика, и учитель решил спросить одного из них. Какое минимально возможное количество вопросов нам надо задать учителю, чтобы наверняка определить, кого именно он решил спросить?*

Закон аддитивности. Алфавитный подход к измерению информации.
Урок 3.

Пусть нам теперь необходимо отгадать сразу два независимых предмета x_1 и x_2, про которые известно, что x_1 принадлежит множеству X_1, содержащему N_1 элементов, а x_2 принадлежит множеству X_2, содержащему N_2 элементов. Вполне допустимо считать, что мы должны угадать пару (x_1, x_2) во множестве X всех возможных пар (x_1, x_2), где $x_1 \in X_1$, а $x_2 \in X_2$. Тогда по формуле Хартли для угадывания задуманной пары необходимо задать $\log_2 N_1 N_2$ вопросов, т. е. получить $\log_2 N_1 N_2$ бит информации. Вместе с тем, элементы x_1 и x_2 можно угадывать независимо. Для угадывания x_1 нам понадобится $\log_2 N_1$ вопросов, а для угадывания x_2 — $\log_2 N_2$. Всего при этом понадобится $\log_2 N_1 + \log_2 N_2$ вопросов (бит информации).

Закон аддитивности. Алфавитный подход к измерению информации.

- Согласно основному логарифмическому тождеству:

$$\log_2 N_1 N_2 = \log_2 N_1 + \log_2 N_2.$$

Закон аддитивности информации:

Количество информации необходимое для установления пары (x1, x2), равно сумме количеств информации необходимых для независимого установления элементов x1 и x2.

ПРИМЕР 4 с. 267

Количество информации, содержащееся с одном символе, называется его ***информационным весом*** и рассчитывается по формуле Хартли $\log_2 N.$

Закон аддитивности. Алфавитный подход к измерению информации.

- Согласно основному логарифмическому тождеству:

$$\log_2 N_1 N_2 = \log_2 N_1 + \log_2 N_2.$$

Закон аддитивности информации:

Количество информации необходимое для установления пары (х1, х2), равно сумме количеств информации необходимых для независимого установления элементов х1 и х2.

ПРИМЕР 4 с. 267

Количество информации, содержащееся с одном символе, называется его **информационным весом** и рассчитывается по формуле Хартли $\log_2 N.$

Закон аддитивности. Алфавитный подход к измерению информации.

- Согласно закону аддитивности, количество информации, содержащееся в сообщении, состоящем из ***m*** символов одного и того же алфавита, равно

$$m\log_2 N.$$

ПРИМЕР 5, 6 стр. 268

Дома: §5.4 прочитать, №1,6 с. 269

ОСНОВЫ ТЕОРИИ ИНФОРМАЦИИ

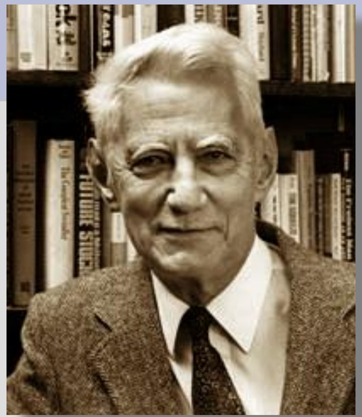
Оптимальное кодирование информации и её сложность.

Урок 7.

Оптимальное кодирование информации и её сложность.

Задача** современной информатики — кодирование информации наиболее* ***коротким *образом*.

Формула Шеннона 1948 г.



$$H = p_1\log_2(1/p_1) + p_2\log_2(1/p_2) + \ldots + p_N\log_2(1/p_N).$$

(показывает средний информационный вес символов того или иного алфавита или энтропию распределения частот появления символов)

- Не существует универсального способа кодирования произвольного файла с частотными характеристиками встречающихся символов, который бы обеспечивал сжатие до величины **меньшей, чем Н.**

- Алгоритм **RLE**
- Алгоритм **Хафмана**
- **Арифметическое кодирование**

Оптимальное кодирование информации и её сложность.

Построение одного из универсальных алгоритмов кодирования

Различные символы встречаются в тексте с различной частотой, то естественно кодировать их так, чтобы те которые встречаются чаще кодировались более коротко, а другие — длиннее (неравномерный код)

ПРОБЛЕМА. *Как понять, где кончился код одного символа и начался другой?*

1 способ

{длина кода, код}{длина кода, код}...{длина кода, код}

- Код Rice и Дельта-код

2 способ

Префиксный код

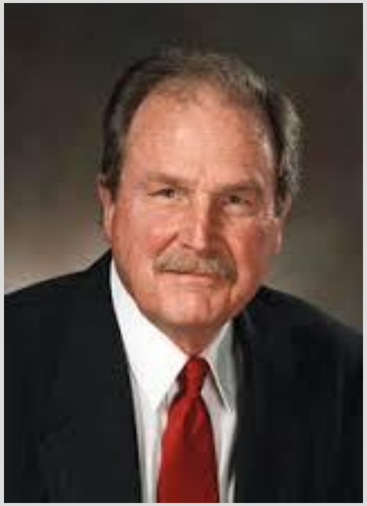
(код одного символа не может быть началом кода другого символа)

Пример 9. Пусть исходный файл состоит только из символов A, B, C и D. При этом имеется 64 буквы A, 32 буквы B, 16 букв C и 16 букв D. Поскольку алфавит в данном случае состоит из четырех символов, каждый символ можно было бы закодировать двумя битами и весь файл поместить в 256 бит. Если же мы закодируем символ A нулевым битом, B — последовательностью битов 10, а символы C и D соответственно последовательностями 110 и 111, то файл сожмется до $64 + 2 \cdot 32 + 3 \cdot 16 + 3 \cdot 16 = 224$ бит. Правда, нам дополнительно придется где-то хранить кодовую таблицу, указывающую, какие символы имеют какие коды.

В данном случае предложенный код 0, 10, 110, 111 является префиксным. Более того, для любого четырехсимвольного алфавита с частотами встречаемости 1/2, 1/4, 1/8, 1/8 рассмотренный способ кодирования является оптимальным, так как каждый символ кодируется двоичной последовательностью длины, совпадающей с информационным весом каждого символа в отдельности. То есть в данном случае достигнута нижняя теоретическая граница сжатия. □

Оптимальное кодирование информации и её сложность.

1. Префиксный код Хаффмана



Дэвид Хаффман
(1925-1999)

Наиболее применимый алгоритм построения префиксного кода для произвольного алфавита

(Доказательство на с. 278-279 пособия)

Пример 10. Построим код Хаффмана для алфавита, состоящего из пяти символов a, b, c, d, e с частотами 0,37(a), 0,22(b), 0,16(c), 0,14(d), 0,11(e). Отождествляя d и e, получаем 0,37(a), 0,25(de), 0,22(b), 0,16(c). Объединяя b и c, имеем 0,38(bc), 0,37(a), 0,25(de). Затем 0,62(ade), 0,38(bc). Сопоставим ade код 0, bc — код 1. Возвращаясь к предпоследнему шагу, расщепим символ ade на a и de с кодами 00 и 01. Затем символ bc расщепляется на b и c с кодами 10 и 11. Наконец, расщепив de, получим для исходного алфавита следующие коды: 00(a), 10(b), 11(c), 010(d), 011(e). □

0.37(a),0.22(b), 0.16(c), 0.14(d), 0.11(e)	00(a), 10(b), 11(c), 010(d), 011(e)
0.37(a), 0.25(de), 0.22(b), 0.16(c)	00(a),01(de),10(b),11(c)
0.37(a), 0.38(bc), 0.25(de),	1(bc), 00(a),01(de)
0.62(ade), 0.38(bc)	0(ade),1(bc)

Оптимальное кодирование информации и её сложность.

2. Сжатие файлов на основе **анализа сложности (RLE, LZ)**

Определение 10. *Сложность объекта или явления* (по Колмогорову) — это минимальное число двоичных знаков (например, нулей и единиц), последовательностью которых можно описать (закодировать) всю информацию об объекте (явлении), достаточное для его дальнейшего воспроизведения (декодирования).

Например: последовательность 01010101010101 является менее сложной, чем 1001110000, т. к. первую можно заменить на более короткую, построенную по другим правилам: $7(01)=111_2(01)$, а пути сокращения второй — не очевидны.

Является ли число π =3,14159256... сложным?

π = длина любой окружности / её диаметр

Оптимальное кодирование информации и её сложность.

Практическая работа «Префиксный код Хаффмана»

- Задание № 2 с. 280

а) построить код Хаффмана

б) оценить размер полученного файла(см. пример 9) и энтропию распределения (ф. Шеннона)

Дома: п. 5.6, з №1,3. Подготовится к к.р.