Меры информации

Информатика

Классификация мер информации

Синтаксическая мера информации

Семантическая мера информации

Прагматическая мера информации

Единицы измерения информации

1 бит = кол-во двоичных цифр (0 и 1)

*Пример: код **11001011** имеет объем данных **V= 8 бит***

1 байт = 8 бит

1 Кбайт = 1024 байт = 2^{10} байт

1 Мбайт = 1024 Кбайт = 2^{20} байт =1 048 576 байт;

1 Гбайт = 1024 Мбайт = 2^{30} байт = 1 073 741 824 байт;

1 Тбайт = 1024 Гбайт = 2^{40} байт = 1 099 511 627 776 байт.

Вероятностный подход

События, о которых нельзя сказать произойдут они или нет, пока не будет осуществлен эксперимент, называются ***случайными.***

Отдельный повтор случайного события называется ***опытом***, а интересующий нас исход этого опыта – ***благоприятным***.

Если ***N*** – общее число опытов, а N_A - количество благоприятных исходов случайного события ***A***, то отношение N_A / N, называется ***относительной частотой появления события A***.

В разных сериях опытов частота может быть различна, но при увеличении количества опытов относительная частота все меньше отклоняется от некоторой константы, ее наличие называется ***статической устойчивостью частот***.

Если все исходы опыта конечны и равновозможные, то их вероятность равна

$$P = \frac{1}{n}$$

где **n** - число исходов.

Энтропия (часть1)

Энтропия – численная мера измеряющая неопределенность.

$$H = f(n)$$

Некоторые свойства функции:

1. ***f (1)=0***, так как при **n=1** исход не является случайным и неопределенность отсутствует.
2. ***f (n)*** возрастает с ростом **n**, чем больше возможных исходов, тем труднее предсказать результат.
3. ***Если a u b*** два независимых опыта с количеством равновероятных исходов n_a и n_b , то мера их суммарной неопределенности равна сумме мер неопределенности каждого из опытов:

$$f(n_a) + f(n_b) = f(n_a, n_b)$$

За количество информации - разность неопределенностей “ДО” и ” ПОСЛЕ” опыта:

$$I = H1 - H2$$

5

Энтропия (часть 2)

$$X = N^M$$

общее число исходов

M – число попыток (пример: **X = 6^2 = 36)**

Энтропия системы из M бросаний кости будет в M раз больше, чем энтропия системы однократного бросания кости - **принцип аддитивности энтропии:**

$$f(N^M) = M \cdot f(N)$$

$$\ln X = M \cdot \ln N \Rightarrow M = \frac{\ln X}{\ln N}$$

$$f(X) = \frac{\ln X}{\ln N} \cdot f(N)$$

Формула Хартли и Шеннона

Обозначим через ***K***

$$K = \frac{f(N)}{\ln N} = \frac{1}{\ln 2}$$

Получим ***f(X) = K · lnX*** или ***H = K · lnX***, таким образом получим формулу **Хартли** для равновозможных исходов

$$H = \log_2 N$$

Формула **Шеннона** для неравновозможных исходов

$$H = \sum_{i=1}^{N} P_i \cdot \log_2 \left(\frac{1}{p_i} \right)$$

Сопоставление мер информации

Мера информации	Единицы измерения	Примеры (для компьютерной области)
Синтаксическая: шенноновский подход	Степень уменьшения неопределенности	Вероятность события
компьютерный подход	Единицы представления информации	Бит, байт. Кбайт и та
Семантическая	Тезаурус	Пакет прикладных программ, персональный компьютер, компьютерные сети и т.д.
	Экономические показатели	Рентабельность, производительность, коэффициент амортизации и тд.
Прагматическая	Ценность использования	Емкость памяти, производительность компьютера, скорость передачи данных и т.д.
		Денежное выражение
		Время обработки информации и принятия решений

Кодирование информации.

Информатика

Абстрактный алфавит

Алфавит - множество знаков, в котором определен их порядок (общеизвестен порядок знаков в русском алфавите: А, Б,..., Я)

1. Алфавит прописных русских букв
2. Алфавит Морзе
3. Алфавит клавиатурных символов ПЭВМ IBM (русифицированная клавиатура)
4. Алфавит знаков правильной шестигранной игральной кости
5. Алфавит арабских цифр
6. Алфавит шестнадцатиричных цифр
7. Алфавит двоичных цифр
8. Двоичный алфавит «точка, «тире»
9. Двоичный алфавит «плюс», «минус»
10. Алфавит прописных латинских букв
11. Алфавит римской системы счисления
12. Алфавит языка блок-схем изображения алгоритмов
13. Алфавит языка программирования

Математическая постановка задачи кодирования

- A - первичный алфавит. Состоит из N знаков со средней информацией на знак I^A .
- B - вторичный алфавит из M знаков со средней информацией на знак I^B .
- Сообщение в первичном алфавите содержит n знаков, а закодированное – m знаков.
- I_s (A)-информация в исходном сообщении,
  I_f (B)-информация в закодированном сообщении.

Математическая постановка задачи кодирования

- $I_S(A) \le I_f(B)$ – условие обратимости кодирования, т.е не исчезновения информации.

  $n^* I^A \le m^* I^B$ (заменили произведением числа знаков на среднее информационное содержание знака).
- m/n –характеризует среднее число знаков вторичного алфавита, который используется для кодирования одного знака первичного. Обозначим его К (А, В)
- $К(A, B) \ge I(A) / I(B)$ Обычно $К(A, B) > 1$
- $К^{min}(A, B) = I^{(A)} / I^{(B)}$ – минимальная длинна кода

12

Первая теорема Шеннона

Примером избыточности может служить предложение
«в словох всо глосноо зомононо боквой о»

Существует возможность создания системы эффективного кодирования дискретных сообщений, у которой среднее число двоичных символов на один символ сообщения асимптотически стремится к энтропии источника сообщений .

X* = {x*i* } - кодирующее устройство – *B

Требуется оценить минимальную среднюю длину кодовой комбинации.

$$n_{ср} = n_i P_i(среднее)$$

Шенноном была рассмотрена ситуация, когда при кодировании сообщения в первичном алфавите учитывается различная вероятность появления знаков, а также равная вероятность появления знаков вторичного алфавита.

Тогда:

$$K^{\min}(A,B) = \frac{I^{(A)}}{\log_2 M} = I^{(A)}$$

где I (A) - средняя информация на знак первичного алфавита.

Вторая теорема Шеннона

При наличии помех в канале всегда можно найти такую систему кодирования, при которой сообщения будут переданы с заданной достоверностью. При наличии ограничения пропускная способность канала должна превышать производительность источника сообщений.

1. Первоначально последовательность $X = \{x_i\}$ кодируется символами из B так, что достигается максимальная пропускная способность (канал не имеет помех).
2. Затем в последовательность из B длины n вводится r символов и по каналу передается новая последовательность из $n + r$ символов. Число возможных последовательностей длины $n + r$ больше числа возможных последовательностей длины n. Множество всех последовательностей длины $n + r$ может быть разбито на n подмножеств, каждому из которых сопоставлена одна из последовательностей длины n. При наличии помехи на последовательность из $n + r$ символов выводит ее из соответствующего подмножества с вероятностью сколь угодно малой.

Это позволяет определять на приемной стороне канала, какому подмножеству принадлежит искаженная помехами принятая последовательность длины $n + r$, и тем самым восстановить исходную последовательность длины n.

14

Вторая теорема Шеннона

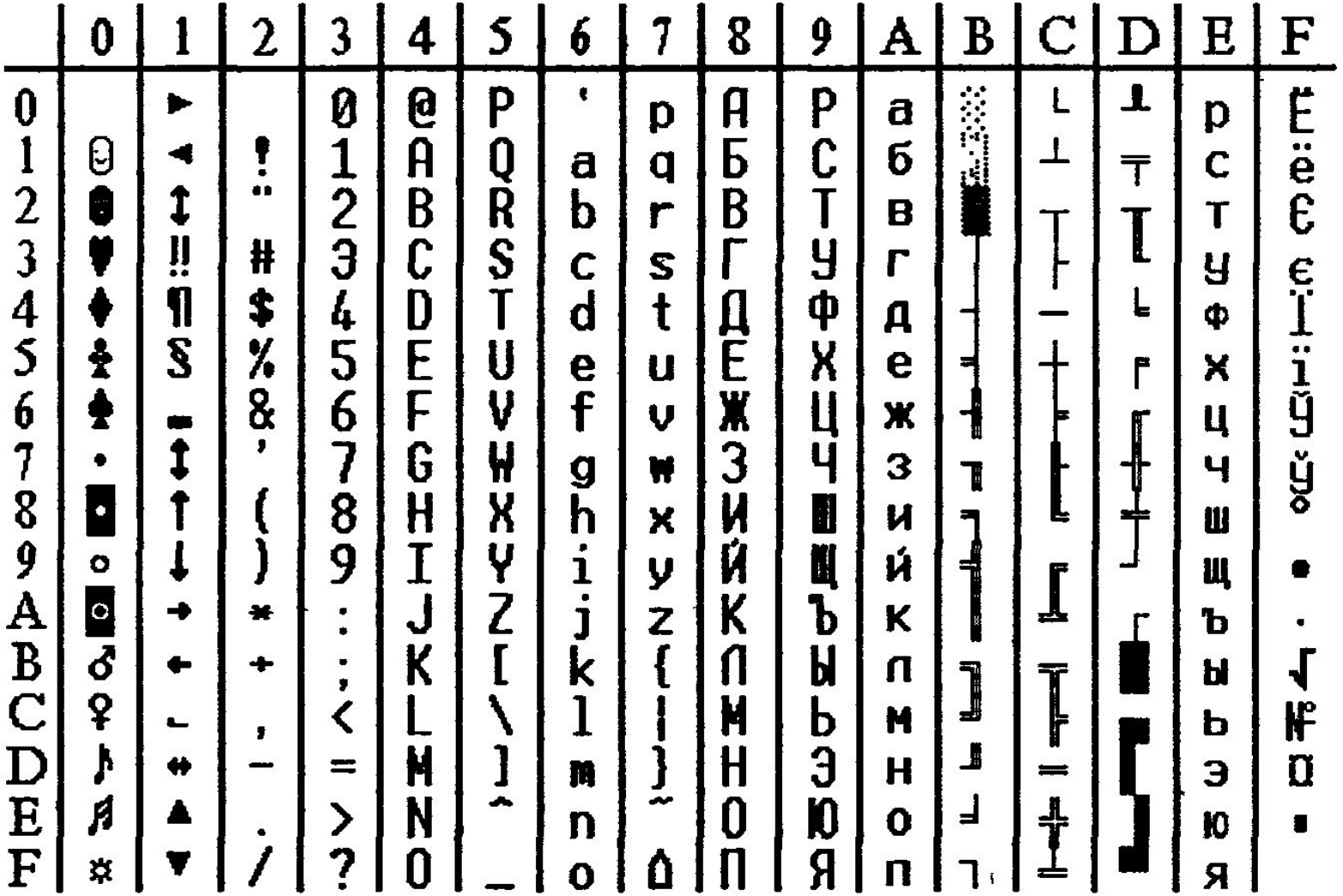
Это позволяет определять на приемной стороне канала, какому подмножеству принадлежит искаженная помехами принятая последовательность длины $n + r$, и тем самым восстановить исходную последовательность длины n.

Эта теорема не дает конкретного метода построения кода, но указывает на пределы достижимого в создании помехоустойчивых кодов, стимулирует поиск новых путей решения этой проблемы.

1. Способ кодирования только устанавливает факт искажения сообщения, что позволяет потребовать повторную передачу.
2. Используемый код находит и автоматически исправляет ошибку передачи.

15

Таблица кодировки ASCII



16