Меры информации

Информатика

Наука исследования свойств знаков и систем

Семиотика (греч. semeion – знак, признак) - наука, занимающаяся исследованием свойств знаков и знаковых систем.

Семиотика выделяет следующие уровни передачи информации:

- ***Синтаксический***, рассматриваются внутренние свойства сообщений.
- ***Семантический***, анализируется смысловое содержание сообщения, его отношение к источнику информации
- ***Прагматический***, рассматривается потребительское содержание сообщения, его отношение к получателю.

Синтаксический уровень

Идея: Это технические проблемы совершенствования методов передачи сообщений и их материальных носителей - сигналов. Проблемы доставки получателю сообщений. Полностью абстрагируются от смыслового содержания сообщений и их целевого предназначения

Информация - данные

Учитывают:

- Тип носителя
- Способ представления информации
- Скорость передачи и обработки
- Размеры кодов представления информации и т.д.

Семантический уровень

Идея: Проблемы связаны с формализацией и учетом смысла передаваемой информации. Проблемы этого уровня чрезвычайно сложны, так как смысловое содержание информации больше зависит от получателя, чем от семантики сообщения, представленного на каком-либо языке.

На данном уровне:

- Анализируется сведенья, которые отражает информация
- Выявляется смысл информации
- Выявляется содержание информации
- Осуществляется обобщение

Прагматический уровень

Идея: Проблемы этого уровня связаны с определением ценности и полезности информации для потребителя. Интересуют последствия от получения и использования данной информации потребителем.

На данном уровне:

- Ценность информации может быть различной для разных потребителей.
- Фактор доставки актуальности доставки и использования.

Классификация мер информации

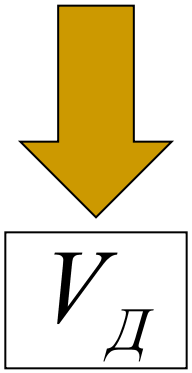
Синтаксическая мера информации

Семантическая мера информации

Прагматическая мера информации

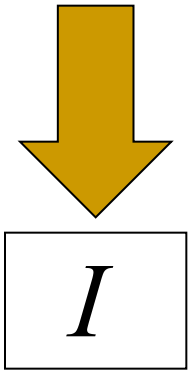
Синтаксическая мера информации

На синтаксическом уровне для измерения информации вводятся два параметра



$V_д$

Объем информации (объемный подход)



I

Количество информации (вероятностный подход)

Объемный подход ($V_д$)

Идея: Если количество информации, содержащейся в сообщении из одного символа, принять за единицу, то объем информации (данных) $V_д$ в любом другом сообщении будет равен количеству символов (разрядов) в этом сообщении. В памяти компьютера объем информации записывается двоичными знаками и равен количеству требуемых для этой записи двоичных кодов.

Объём данных ($V_д$) в техническом смысле этого слова как информационный объём сообщения или как объём памяти, необходимый для хранения сообщения без каких-либо изменений.

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:

$$I = H1 - H2$$

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 с учетом (1): ***H = K · lnN***, таким образом получим формулу **Хартли** для равновозможных исходов

$$H = \log_2 N$$

Формула **Шеннона** для неравновозможных исходов

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

Энтропия (часть 3)

Информация – это содержание сообщения, понижающего неопределенность некоторого опыта с неоднозначным исходом; убыль связанной с ним энтропии является количественной мерой информации.

Количество информации (в битах), заключенное в двоичном слове, равно числу двоичных знаков в нем.

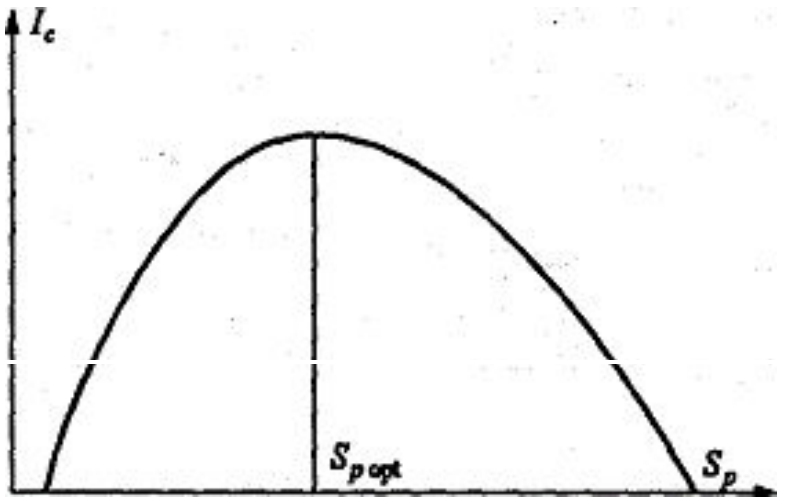
$$H = \log_2 2 = 1 бит$$

Семантическая мера информации

Тезаурус — это совокупность сведений, которыми располагает пользователь или система.

при Sp→0 пользователь не воспринимает, не понимает поступающую информацию;

при Sp→∞ пользователь все знает, и поступающая информация ему не нужна.



Зависимость количества семантической информации, воспринимаемой потребителем, от его тезауруса.

$$C = \frac{I}{V_Д}$$

количество семантической информации

Прагматическая мера информации

Эта мера определяет полезность информации (ценность) для достижения пользователем поставленной цели. Эта мера также величина относительная, обусловленная особенностями использования этой информации в той или иной системе.

Ценность информации целесообразно измерять в тех же самых единицах (или близких к ним), в которых измеряется **целевая функция**.

Целевая функция - есть математическое выражение некоторого критерия качества одного объекта (решения, процесса и т.д.) в сравнении с другим.

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

Кодирование информации.

Источник представляет сообщение в алфавите, который называется ***первичным***, далее это сообщение попадает в устройство, преобразующее и представляющее его во ***вторичном алфавите***.

Код – правило, описывающее соответствие знаков (или их сочетаний) первичного алфавита знаком (их сочетаниями) вторичного алфавита.
Кодирование – перевод информации, представленной сообщением в первичном алфавите, в последовательность кодов.
Декодирование – операция обратная кодированию.
Кодер – устройство, обеспечивающее выполнение операции кодирования.
Декодер – устройство, производящее декодирование.

Операции кодирования и декодирования называются ***обратимыми,*** если их последовательное применение обеспечит возврат к исходной информации без каких-либо ее потерь.

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

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 $X = \{xi\}$ 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 + m 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.

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.

Вторая теорема Шеннона

1) способ кодирования только устанавливает факт искажения сообщения, что позволяет потребовать повторную передачу;

2) используемый код находит и автоматически исправляет ошибку передачи.

Особенности вторичного алфавита при кодировании

1. Элементарные коды 0 и 1 могут иметь одинаковые длительности (t0 = t1) или разные (≠).

2. Длина кода может быть одинаковой для всех знаков первичного алфавита (***код равномерный***) или различной (***неравномерный код***)

3. Коды могут строиться для отдельного знака первичного алфавита (***алфавитное кодирование***) или для их комбинаций (***кодирование блоков, слов***).

Равномерное алфавитное кодирование. Представление чисел в компьютере

1. Компьютерный алфавит С включает буквы латинского алфавита – 52 шт.

2. Букв русского (прописные и строчные) – 66 шт.

3. Цифры 0…9 – 10 шт.

4. Знаки математических операций, препинания, спецсимволы – 20 штук

Итого: 148

$$К\ (С,\ 2) \geq \log_2 148 \geq 7{,}21,$$

так как длина кода – целое число, следовательно,

$$К\ (С,2) = 8бит = 1байт.$$

Именно такой способ кодирования принят в компьютерных системах. Один байт соответствует количеству информации в одном знаке алфавита при их равновероятном распределении.

Это объемный способ измерения информации.

Присвоение символу конкретного двоичного кода фиксируется в кодовой таблице, где устанавливается соответствие между символами и их порядковыми номерами.

Таблицы кодировки

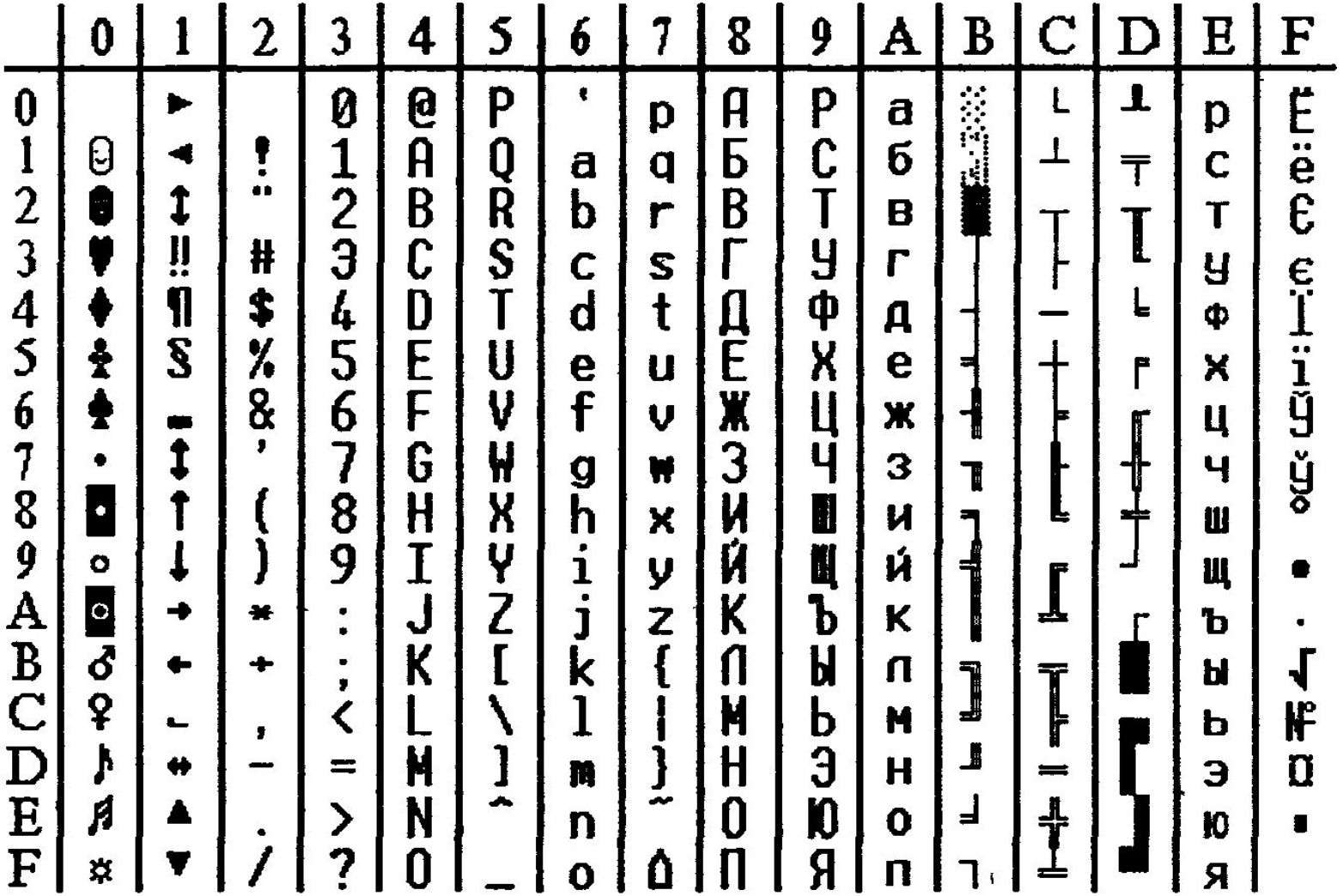
Таблица, в которой устанавливается однозначное соответствие между символами и их порядковыми номерами, называется **таблицей кодировки**.

Для разных типов ЭВМ используют различные таблицы кодировки:

ANSI - (American National Standards Institute)

ASCII - (American Standard Cod for Information Interchange)

Таблица кодировки ASC II



Код обмена информации ASC II

Первоначально – 7 бит

$N=2^7=128$ символов

0…31- всевозможные управляющие символы

32…127 – видимые на экране символы.

Сейчас – 8 бит

$N=2^8=256$ символов

128…255- национальные алфавиты,псевдографика

01000001 = буква А = 65

Системы кодирования

КОИ-7

Windows-1251

КОИ-8

ISO

Unicode

Кодирование текстовой информации

$$I = \log_2 256 = 8 \qquad = 1$$

$$V = 8N$$

Кодирование графической информации*

Растровое изображение представляет собой однослойную сетку точек, называемых пикселами (pixel, от англ. picture element). Код пиксела содержит информации о его цвете.

Векторное изображение многослойно. Каждый элемент векторного изображения - линия, прямоугольник, окружность или фрагмент текста - располагается в своем собственном слое, пикселы которого устанавливаются независимо от других слоев.

Объем графического файла в битах определяется как произведение количества пикселей на разрядность цвета (битовую глубину)

$$N * M$$

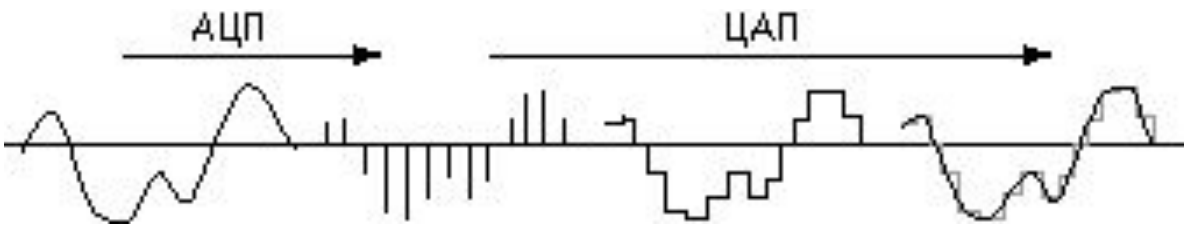
Кодирование графической информации

Бит/пиксель	4 бита	8 бит	16 бит	24 бита
Число цветов	2^4=16цв	2^8=256цв	2^{16}= 65536цв	2^{24}= 16777216цв
640x480	150 Кбайт	300 Кбайт	600 Кбайт	900 Кбайт
800x600	234,4 Кбайт	468,8 Кбайт	937,6 Кбайт	1,4 Мбайт
1024x768	384 Кбайт	768 Кбайт	1,5 Мбайт	2,25 Мбайт
1280x1024	640 Кбайт	1,25 Мбайт	2,5 Мбайт	3,75 Мбайт

Кодирование звука*

Звук - это колебания воздуха

Процесс преобразования аналогового сигнала в последовательность двоичных чисел называется **дискретизацией (или оцифровкой)**, а устройство, выполняющее его - **аналого-цифровым преобразователем (АЦП).**



Для того чтобы воспроизвести закодированный таким образом звук, нужно выполнить обратное преобразование (для него служит *цифро-аналоговый преобразователь* -- ЦАП), а затем сгладить получившийся ступенчатый сигнал.

Кодирование видеоинформации*

Число кадров вычисляется как произведение длительности видеоклипа на скорость кадров , то есть их количество в 1 с

$$V = N * M * C * v * \Delta t$$

При разрешении 800*600 точек, разрядности цвета С=16, скорости кадров v=25 кадров/с, видеоклип длительностью 30 с будет иметь объем:

V=800*600*16*25*30=576*107(бит)=72*107(байт)=687(Мбайт)

Диплом - документ, подтверждающий, что у вас была потенциальная возможность чему-то научиться

Арон Вигушин