По материалам сайтов
http://marknet.narod.ru/spr/list5.htm
http://n-t.ru/ri/ch/pi02.htm

ИЗМЕРЕНИЕ ИНФОРМАЦИИ

Подходы к изучению информации:
- аксиологический (с точки зрения содержания);
- прагматический (с точки зрения необходимости и достаточности информации для принятия решения);
- семантический (связь информации с ранее накопленной системой знаний);
- структурный (представление, организация, способ хранения);
- статистический (количественные меры).

Из перечисленных подходов нас интересуют последние 3.

Семантический подход к изучению информации

Данный подход трудно формализуем, и пока что точных способов измерения смыслового содержания информации нет. Наибольшее применение получила так называемая тезаурусная мера.

Тезаурус – полный набор накопленных сведений о предмете информации.

Пусть S – состояние тезауруса потребителя информации.

Получена некоторая информация I.
Ясно, что смысловое содержание этой информации I_c (информация, которая может пополнить тезаурус, т.е. понятная получателю информации и в то же время новая для него) зависит от исходного состояния тезауруса, т.е. $I_c=f(S)$. Действительно, если ваш тезаурус в области, например, математики близок к нулю, и вы незнакомы с понятием «степень», вы ничего не поймёте из объяснений учителя о том, что такое «логарифм». С другой стороны, если вы закончили Матмех, вряд ли вы узнаете что-то новое для себя на школьном уроке математики. Интуитивно ясно, что между этими крайними ситуациями должен существовать некий оптимум, т.е. такое состояние тезауруса, при котором из информации будет извлечено больше всего смысла. Разработчики учебных программ, рекламных материалов, журналисты как раз и занимаются поиском такого оптимума. Но для нас как программистов это не очень актуально – разве что при разработке документации.

Структурный подход к изучению информации

Общеизвестно: от того, как организована информация, зависит, насколько удобно с ней манипулировать. Сплошной текст воспринимается хуже, чем разбитый на фрагменты с заголовками. Если заметки написаны на отдельных бумажках – их легче удалять и добавлять, чем если они на большом листе, но хранить и передавать труднее. Если информация о товарах упорядочена по названиям фирм – легко найти шиповки Iceburg, но трудно найти шиповки вообще… Примеров – множество, есть некоторые общие правила и конкретные приёмы структуризации

информации – но о них в других частях данного курса и в других курсах (например, БД).

Статистический подход к измерению информации

1) Вероятностный анализ

N – количество возможных исходов какого-то события (количество возможных значений чего-либо, например, задуманного числа).

H – энтропия – величина неопределённости, связанной с исходом события.

H=f(N) – энтропия зависит от числа возможных исходов.

Это – возрастающая функция, так как с ростом N растёт неопределённость.

Пусть получена некоторая информация об исходе события.

H_1 – неопределённость до получения информации.

H_2 – неопределённость после получения информации.

$H_1 \geq H_2$ (с получением информации неопределённость не возрастает, она может остаться неизменной, если информация бесполезна).

I – количество информации в сообщении.

$I = H_1 - H_2$

Остаётся придумать, как считать неопределённость.

Верно ли, что неопределённость прямо пропорциональна количеству исходов? Иными словами, верно ли, что если задумано одно из 8 чисел, его в 2 раза труднее отгадать, чем задуманное из 4 чисел?

Предположим, что числа отгадывают, задавая вопросы, на которые можно получить ответ ДА или НЕТ, например: «Задуманное число меньше 5?». Сколько вопросов потребуется задать, чтобы отгадать одно из 8 задуманных чисел? Если действовать самым рациональным путём, каждым вопросом отсекая от множества возможных значений половину – то 3. А из 4? А из 16? Выразимколичество требуемых вопросов через N.

$H = \log_2 N$ – формула Хартли

Формула применима только тогда, когда все исходы можно считать равновероятными.

Очевидно, H будет равно единице при N = 2. Иначе говоря, в качестве единицы принимается количество информации, связанное с проведением опыта, состоящего в получении одного из двух равновероятных исходов (примером такого опыта может служить бросание монеты при котором возможны два исхода: «орел», «решка»). Такая единица количества информации называется «**бит**» - от BInary digiT, «двоичная цифра».

Пример

Бросили кубик с 6 гранями. Имеются три высказывания о результате броска:
1) выпало чётное число;
2) выпало число, меньшее 3;

3) выпало не число 5.
Определить количество информации в каждом высказывании.

До получения информации:
N=6. $H = \log_2 6 \approx 2.59$

После получения высказывания 1:
N=3 $H_1 = \log_2 3 \approx 1.59$ $I_1 = 2.59 - 1.59 = 1$

После получения высказывания 2:
N=2 $H_2 = \log_2 2 = 1$ $I_1 = 2.59 - 1 = 1.59$

После получения высказывания 1:
N=5 $H_3 = \log_2 5 \approx 2.33$ $I_1 = 2.59 - 2.33 = 0.26$

Вернёмся к примечанию о равновероятных событиях. А если они не равновероятны?

Ситуация: детишки отгадывают задуманные слова, уже известно, что задумано слово на С, и вот поступает информация о второй букве. Одинаковую ли информацию мы получим, если узнаем, что «вторая буква - О» и «Вторая буква – Ю»?

Предположим, что есть N возможным исходов (N возможных вариантов). В нашей задаче это количество слов на С с заданной второй буквой. Предположим, на СО начинается 200 слов, а на СЮ – одно. Ну а количество слов на С, для ровного счёта, 4000. Ясно, что при открытии О неопределённость сократилась в 20 раз (примерно 4,5 бита мы получили), а при открытии Ю в 4000 раз (почти 12 бит).

А как посчитать, сколько бит в среднем несёт в себе вторая буква слова на букву С? Пусть есть 33 варианта второй буквы (i=1…33). Для каждого i мы знаем Ni – количество слов с таким началом. По формуле Хартли $Hi = \log_2 Ni$. Но вероятность. Что откроется именно i-я буква (Pi) для разных букв разная, причём Pi=1 / Ni (по классическому определению вероятности – «отношения числа благоприятных исходов к числу возможных»). Тогда среднее количество информации во второй букве слова на С равняется

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

Это – **формула Шеннона.**

Несложно доказать,, что при равновероятных событиях формула Шеннона сводится к формуле Хартли.

Частая ошибка: как только появляется в задаче что-то. Напоминающее неравновероятные исходы, вспоминают формулу Шеннона… и оказываются неправы! Эта формула позволяет вычислять «средние» количества информации. Она представляет ценность скорее с точки зрения прагматического подхода к изучению информации (например, какую букву слова выгоднее открыть, чтобы угадать его – первую или последнюю?). А в большинстве задач удобнее исходить из формулы Хартли и общей формулы информации как разности энтропий. А

большинство задач вообще решаются из определения: 1бит – информация, уменьшающая неопределённость вдвое.

Пример

) Сельский клуб, дискотека.
- Клава, а ты где живёшь?
- В Ромашкино.
- Это такая маленькая деревня, за мостом?
- Ага. Там всего 32 дома. И 2 улицы – Продольная и Поперечная.

- А ты на какой живёшь?
- На Поперечной.

Сколько домов на Поперечной улице, если в последнем сообщении Клавы содержится 4 бита информации об её точном адресе?

/Начальная неопределённость 5 бит, 4 убрали, остался 1, итого 2 дома. А можно через определение бита: получили 4 бита – уменьшили количество возможностей в 16 раз, 32:16=2/

2) Объёмный анализ

Измерение информации можно основывать на том, сколько двоичных разрядов потребуется для её кодирования. А как кодировать?

Самый простой код – просто пронумеровать и перевести в двоичную систему. Сколько разрядов понадобится для кодирования чего-либо, имеющего N возможных значений?

$\log_2 N$, округлённый в бОльшую сторону. Ой, опять формула Хартли. Случайно ли это? Представьте себе, что мы отгадываем число от 0 до 7, задавая «правильные» вопросы, разбивающие множество возможных значений пополам, и при ответе ДА пишем 1, при НЕТ – 0. Задумали 5.

Число больше 3? ДА. Пишем 1.

Число больше 5? НЕТ. Пишем 0.

Число больше 4? ДА. Пишем 1. Отгадали. А записана у нас как раз двоичная запись числа 5.

Но здесь у нас все исходы равновероятны. Потому мы и применяем такой простой код. А если нам надо кодировать нечто, имеющее 18 различных значений, причём первые 2 значения встречаются с вероятностью 1/4, а остальные 16 с вероятностью 1/32? Можно тупо воспользоваться тем же приёмом – пронумеровать. Тогда придётся отвести под этот показатель (сколько?) 5 разрядов. А можно иначе. Первые два закодируем как 00 и 01, а остальные закодируем 4-разрядными числами от 1000 до 1111. Код раскодируется однозначно: по первой цифре определим длину. Сколько в среднем разрядов будет приходиться на 1 слово? В половине случаев 2, в остальных случаях 4, в среднем 3. Хорошо ли мы закодировали? Проверим по формуле Шеннона:

$2*(1/4 * \log_2 4) + 16*(1/32*\log_2 32)=3.5$ бита. Наш код дал результат лучше!

Мораль: формулы для измерения информации реально связаны с практикой программирования, а не просто страшилки для ЕГЭ.