Практическое занятие №2
Использование формулы Хартли при решении задач на определение количества информации

Цель работы: экспериментальное изучение количественных аспектов информации.

Методические указания.

Понятие количество информации отождествляется с понятием информация. Эти два понятия являются синонимами. Мера информации должна монотонно возрастать с увеличением длительности сообщения (сигнала), которую естественно измерять числом символов в дискретном сообщении и временем передачи в непрерывном случае. Кроме того, на содержание количества информации должны влиять и статистические характеристики, так как сигнал должен рассматриваться как случайный процесс.

При этом наложено ряд ограничений:

1. Рассматриваются только дискретные сообщения.
2. Множество различных сообщений конечно.
3. Символы, составляющие сообщения равновероятны и независимы.

Хартли впервые предложил в качестве меры количества информации принять логарифм числа возможных последовательностей символов.

$$I=\log m^k=\log N \qquad (1)$$

К.Шеннон попытался снять те ограничения, которые наложил Хартли. На самом деле в рассмотренном выше случае равной вероятности и независимости символов при любом k все возможные сообщения оказываются также равновероятными, вероятность каждого из таких сообщений равна P=1/N. Тогда количество информации можно выразить через вероятности появления сообщений I=–log P.

В силу статистической независимости символов, вероятность сообщения длиной в k символов равна

$$P=\prod_{i=1}^{k} p_i$$

Если i–й символ повторяется в данном сообщении k_i раз, то

$$P=\prod_{i=1}^{m} p_i^{k_i}$$

так как при повторении i символа k_i раз k уменьшается до m. Из теории вероятностей известно, что, при достаточно длинных сообщениях (большое число символов k) $k_i \approx k \cdot p_i$ и тогда вероятность сообщений будет равняться

$$P=\prod_{i=1}^{m} p_i^{kp_i}$$

Тогда окончательно получим

$$I=-\log P=-k\sum_{i=1}^{m} p_i \log p_i \qquad (2)$$

Данное выражение называется формулой Шеннона для определения количества информации.

Формула Шеннона для количества информации на отдельный символ

сообщения совпадает с энтропией. Тогда количество информации сообщения состоящего из k символов будет равняться I=k·H

Количество информации, как мера снятой неопределенности

При передаче сообщений, о какой либо системе происходит уменьшение неопределенности. Если о системе все известно, то нет смысла посылать сообщение. Количество информации измеряют уменьшением энтропии.

Количество информации, приобретаемое при полном выяснении состояния некоторой физической системы, равно энтропии этой системы:

$$I = -\sum_{i=1}^{n} p_i \log p_i$$

Количество информации I – есть осредненное значение логарифма вероятности состояния. Тогда каждое отдельное слагаемое $-\log p_i$ необходимо рассматривать как частную информацию, получаемую от отдельного сообщения, то есть

$$I_i = -\log p_i$$

Избыточность информации

Если бы сообщения передавались с помощью равновероятных букв алфавита и между собой статистически независимых, то энтропия таких сообщений была бы максимальной. На самом деле реальные сообщения строятся из не равновероятных букв алфавита с наличием статистических связей между буквами. Поэтому энтропия реальных сообщений $-H_p$, оказывается много меньше оптимальных сообщений – H_o. Допустим, нужно передать сообщение, содержащее количество информации, равное I. Источнику, обладающему энтропией на букву, равной H_p, придется затратить некоторое число n_p, то есть

$$I = n_p H_p$$

Если энтропия источника была бы H_0, то пришлось бы затратить меньше букв на передачу этого же количества информации

$$n_o = \frac{I}{H_o} < n_p$$

$I= n_0H_0$

Таким образом, часть букв $n_p–n_o$ являются как бы лишними, избыточными. Мера удлинения реальных сообщений по сравнению с оптимально закодированными и представляет собой избыточность D.

$$D = 1 - \frac{H_p}{H_o} = 1 - \frac{n_o}{n_p} = \frac{n_p - n_o}{n_p} \qquad (3)$$

Но наличие избыточности нельзя рассматривать как признак несовершенства источника сообщений. Наличие избыточности способствует повышению помехоустойчивости сообщений. Высокая избыточность естественных языков обеспечивает надежное общение между людьми.

Частотные характеристики текстовых сообщений

Важными характеристиками текста являются повторяемость букв, пар букв (биграмм) и вообще m-ок (m-грамм), сочетаемость букв друг с другом, чередование гласных и согласных и некоторые другие. Замечательно, что эти характеристики являются достаточно устойчивыми.

Идея состоит в подсчете чисел вхождений каждой n^m возможных m-грамм в достаточно длинных открытых текстах $T=t_1t_2…t_l$, составленных из букв алфавита $\{a_1, a_2, ..., a_n\}$. При этом просматриваются подряд идущие m-граммы текста

$t_1t_2...t_m$, $t_2t_3...t_{m+1}$, ..., $t_{i-m+1}t_{l-m+2}...t_l$.

Если $\vartheta(a_{i1}a_{i2}...a_{im})$ – число появлений m-граммы $a_{i1}a_{i2}...a_{im}$ в тексте T, а L общее число подсчитанных m-грамм, то опыт показывает, что при достаточно больших L частоты

$$\frac{\vartheta(a_{i1}a_{i2}...a_{im})}{L}$$

для данной m-граммы мало отличаются друг от друга.

В силу этого, относительную частоту считают приближением вероятности P ($a_{i1}a_{i2}...a_{im}$) появления данной m-граммы в случайно выбранном месте текста (такой подход принят при статистическом определении вероятности).

Для русского языка частоты (в порядке убывания) знаков алфавита, в котором отождествлены Е с Ё, Ь с Ъ, а также имеется знак пробела (-) между словами, приведены в таблице 1.

Таблица 1

-	0.175	О	0.090	Е, Ё	0.072	А	0.062
И	0.062	Т	0.053	Н	0.053	С	0.045
Р	0.040	В	0.038	Л	0.035	К	0.028
М	0.026	Д	0.025	П	0.023	У	0.021
Я	0.018	Ы	0.016	З	0.016	Ь, Ъ	0.014
Б	0.014	Г	0.013	Ч	0.012	Й	0.010
Х	0.009	Ж	0.007	Ю	0.006	Ш	0.006
Ц	0.004	Щ	0.003	Э	0.003	Ф	0.002

Некоторая разница значений частот в приводимых в различных источниках таблицах объясняется тем, что частоты существенно зависят не только от длины текста, но и от его характера.

Устойчивыми являются также частотные характеристики биграмм, триграмм и четырехграмм осмысленных текстов.

Задание

1. Определить количество информации (по Хартли), содержащееся в заданном сообщении, при условии, что значениями являются буквы кириллицы.

«Фамилия Имя Отчество» завершил ежегодный съезд эрудированных школьников, мечтающих глубоко проникнуть в тайны физических явлений и химических реакций

2. Построить таблицу распределения частот символов, характерные для заданного сообщения. Производится так называемая частотная селекция, текст сообщения анализируется как поток символов и высчитывается частота встречаемости каждого символа. Сравнить с имеющимися данными в табл 1.

3. На основании полученных данных определить среднее и полное количество информации, содержащееся в заданном сообщении

4. Оценить избыточность сообщения.

5. Перечислите старинные носители информаци.

6. Заполните пустые ячейки цифрами от 1 до 5. Если параметр ярко выражен - 5, либо поставьте 1 - если параметр выражен слабо:

	Емкость	Скорость обмена	Надежность	Стоимость
Дискета				
Винчестер				
Диск CD-R				
Диск DVD-R				
Flash-память				

7. Выскажите свой прогноз о будущем информационных носителей

Контрольные вопросы:

1. Как должна возрастать мера информации?
2. Как выглядит формула Хартли для измерения количества информации?
3. Как выглядит формула Шенона для измерения количества информации?
4. Что называется, количеством информации?
5. Что такое избыточность информации?