Практическая работа **№ 3**

**Вероятностный подход к измерению дискретной и непрерывной информации**

Цель: научиться вычислять вероятности событий (появление символов в сообщении) и рассчитывать энтропию.

Время выполнения: 2 часа

Оборудование: ПК.

Программное обеспечение: операционная система, калькулятор, текстовый редактор.

Теоретические основы

Количество информации по Хартли и Шеннону

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 mk=log N (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

D:\работа\ФТФ\Основы теории передачи инф\теория инф и кодирование\Количественные аспекты  информации_files\1_3ite2.gif

Если i−й символ повторяется в данном сообщении ki раз, то

D:\работа\ФТФ\Основы теории передачи инф\теория инф и кодирование\Количественные аспекты  информации_files\1_3ite3.gif

так как при повторении i символа ki раз k уменьшается до m. Из теории вероятностей известно, что, при достаточно длинных сообщениях (большое число символов k) ki≈k·pi и тогда вероятность сообщений будет равняться

D:\работа\ФТФ\Основы теории передачи инф\теория инф и кодирование\Количественные аспекты  информации_files\1_3ite4.gif

Тогда окончательно получим

D:\работа\ФТФ\Основы теории передачи инф\теория инф и кодирование\Количественные аспекты  информации_files\1_3ite5.gif (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:

D:\работа\ФТФ\Основы теории передачи инф\теория инф и кодирование\Количество информации_files\1_4ite11.gif

Количество информации I − есть осредненное значение логарифма вероятности состояния. Тогда каждое отдельное слагаемое −log pi необходимо рассматривать как частную информацию, получаемую от отдельного сообщения, то есть

D:\работа\ФТФ\Основы теории передачи инф\теория инф и кодирование\Количество информации_files\1_4ite12.gif

## 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 −Hр,оказывается много меньше оптимальных сообщений − Hо. Допустим, нужно передать сообщение, содержащее количество информации, равное I. Источнику, обладающему энтропией на букву, равной Hр, придется затратить некоторое число nр, то есть

D:\работа\ФТФ\Основы теории передачи инф\теория инф и кодирование\Избыточность информации_files\3_2ite81.gif

Если энтропия источника была бы Н0, то пришлось бы затратить меньше букв на передачу этого же количества информации

I= n0H0 D:\работа\ФТФ\Основы теории передачи инф\теория инф и кодирование\Избыточность информации_files\3_2ite82.gif

Таким образом, часть букв nр−nо являются как бы лишними, избыточными. Мера удлинения реальных сообщений по сравнению с оптимально закодированными и представляет собой избыточность D.

D:\работа\ФТФ\Основы теории передачи инф\теория инф и кодирование\Избыточность информации_files\3_2ite83.gif (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 nm возможных m-грамм в достаточно длинных открытых текстах T=t1t2…tl, составленных из букв алфавита {a1, a2, ..., an}. При этом просматриваются подряд идущие m-граммы текста

t1t2...tm, t2t3... tm+1, ..., ti-m+1tl-m+2...tl.

Если mhtml:file://D:\работа\ФТФ\Основы%20теории%20передачи%20инф\Анализ%20текстов.mht!http://www.statsoft.ru/home/portal/exchange/images/textanalysis/im01.gif– число появлений m-граммы ai1ai2...aim в тексте T, а L общее число подсчитанных m-грамм, то опыт показывает, что при достаточно больших L частоты

mhtml:file://D:\работа\ФТФ\Основы%20теории%20передачи%20инф\Анализ%20текстов.mht!http://www.statsoft.ru/home/portal/exchange/images/textanalysis/im02.gif для данной m-граммы мало отличаются друг от друга.

В силу этого, относительную частоту считают приближением вероятности P (ai1ai2...aim) 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 E c 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.

**Порядок выполнения работы**

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

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.

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

1. Построить таблицу распределения частот символов, характерныx для заданного сообщения путём деления количества определённого символа в данном сообщении на общее число символов

По формуле



-

H= вычислил энтропию сообщения

1. Далее по формуле Шеннона для определения кол-ва информации

D:\работа\ФТФ\Основы теории передачи инф\теория инф и кодирование\Количественные аспекты  информации_files\1_3ite5.gif вычислил кол-во информации в передаваемом сообщении

1. Вычислил избыточность D по формуле

D:\работа\ФТФ\Основы теории передачи инф\теория инф и кодирование\Избыточность информации_files\3_2ite83.gif

**Отчет**

Отчет должен быть оформлен в текстовом редакторе и содержать:

* наименование работы;
* цель работы;
* задание;
* последовательность выполнения работы;
* ответы на контрольные вопросы;
* вывод о проделанной работе.

**Контрольные вопросы**

1. Дать определение понятие энтропия?

2. Что означает вероятностный способ измерения информации?

3. Что означает статическое определение вероятности?

4. Запишите уравнение Хартли?

5. Какие основные разработки внес в основу теории информации Шеннон?