Учреждение образования

«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ»

Отчет по лабораторной работе №2

«Элементы теории информации. Параметры и характеристики дискретных информационных систем»

Студентка: Пунько А.А,

ФИТ 3 курс 5 группа

Преподаватель: Берников В. О.

Минск 2020

**Цель:** приобретение практических навыков расчета и анализа параметров и информативных характеристик дискретных ИС

**Задачи:**

1. Закрепить теоретические знания по основам теории информации.
2. Разработать приложение для расчета и анализа параметров и информативных характеристик дискретных ИС.
3. Результаты выполнения лабораторной работы оформить в виде описания разработанного приложения, методики выполнения экспериментов с использованием приложения и результатов эксперимента.
4. **Теоретические сведения**

Передача информации (данных) осуществляется между двумя абонентами, называемыми источником сообщения (ИcС) и получателем сообщения (ПС). Третьим элементом информационной системы является канал (среда) передачи, связывающий ИсС и ПС. Отметим также, что и в системах с хранением информации всегда можно выделить ИcС и ПС. В данном случае каналом передачи здесь выступает устройство хранения информации (память). Например, при записи данных в ОЗУ (оперативное запоминающее устройство) компьютера в качестве ИcС и ПС может выступать процессор (соответственно при записи и чтении данных). Таким образом, простейшая информационная система состоит из трех элементов: источника сообщения, канала передачи сообщения и получателя сообщения. Отображение сообщения обеспечивается изменением какой-либо физической величины, характеризующей процесс (например, амплитуда, частота, фаза). Эта величина является информационным параметром сигнала(в общем случае – информационной системы). Сигналы, как и сообщения, могут быть непрерывными и дискретными. Информационный параметр непрерывного сигнала с течением времени может принимать любые мгновенные значения в определенных пределах. Непрерывный сигнал часто называют аналоговым, а каналы и устройства функционирующие на основе такого типа сигналов – аналоговыми.. Дискретный сигнал (устройство или канал передачи) характеризуется конечным числом значений информационного параметра. Дискретные сообщения состоят из последовательности дискретных знаков. Часто этот параметр принимает всего два значения (0 или 1).

Сообщение или канал его передачи на основе этих двух значений сигнала называют двоичным или бинарным. Построение сигнала по определенным правилам, обеспечивающим соответствие между сообщением и сигналом, называют кодированием. Кодирование в широком смысле – преобразование сообщения в сигнал.

Кодирование в узком смысле – представление исходных знаков, называемых символами, в другом алфавите с меньшим числом знаков. Оно осуществляется с целью повышения надежности и преобразования сигналов к виду, удобному для передачи по каналам связи. Последний тип кодирования относится к так называемой прикладной теории кодирования информации.

Важнейшая характеристика источника, получателя или канала – алфавит. Алфавит А – это общее число знаков или символов (N), используемых генерации или передачи сообщений. Символы алфавита будем обозначать через {аi}, где 1 ≤ i≤ N; N – мощность алфавита. Минимальное число элементов алфавита Nmin=2, А={0,1} – двоичный код. Один дискретный знак представляет собой элементарное сообщение, последовательность знаков – сообщение. Набор элементов алфавита, создаваемых дискретным источником сообщений, заранее, априори (до опыта) известен получателю. ИcC в каждый дискретный момент времени выдает один элемент алфавита. Этот элемент сообщения является одним из символов алфавита. Понятно, что ПС заранее не известно, какой это элемент. Если обозначить вероятность выбора каждого элемента алфавита p(аi), то

|  |  |
| --- | --- |
|  | (1.1) |

Информационной характеристикой алфавита (источника сообщений на основе этого алфавита) является энтропия. Этот термин применительно к техническим системам был введен Шенноном и Хартли. Энтропию алфавита А={ai} по К.Шеннону рассчитывают по следующей формуле:

|  |  |
| --- | --- |
|  | (1.2) |

С физической точки зрения энтропия алфавита показывает, какое количество информации приходится в среднем на один символ алфавита. Частным случаем энтропии Шеннона является энтропия Хартли. Дополнительным условием при этом является то, что все вероятности одинаковы и постоянны для всех символов алфавита. С учетом этого формулу (2.1) можно преобразовать к виду:

|  |  |
| --- | --- |
|  | (1.3) |

Сообщение Хk, которое состоит из k символов, должно характеризоваться определенным количеством информации, I(Хk):

|  |  |
| --- | --- |
| I(Хk) = H(A) \* k | (1.4) |

Здесь Н(А)– энтропия алфавита с соответствующим распределением вероятностей р(аi). Если принять, что р(аi =1)= р(1) и р(аi =0) = р(0), вычислим энтропию бинарного алфавита:

|  |  |
| --- | --- |
| H(A2) = -р(0)\*log2(р(0)) -р(1)\*log2(р(1)) | (1.5) |

К примеру, полагая что сообщение Хk состоит только из единиц (Хk= =11…1) и имеет длину k, т.е. вероятность того, что произвольный символ равен единице, составляет единицу (р(аi =1)=1), и другая вероятность –- р(аi =0)=0 для N i , 1= . Фактически, здесь имеет место использование моноалфавита: алфавита, состоящего из одного символа. Учитывая, что сумма р(1) +р(0)==1 и, выразить одну вероятность через другую (например, р(1) = 1-р(0)), можно теоретически доказать информативность бинарного алфавита, решив дифференциальное уравнение [dH(A)/dp(1)] = 0. Если вероятность ошибки в ДСК отлична от 0 (р> 0), переданное сообщение может содержать ошибки: Хk ≠ Yk. Количество информации в таком сообщении при его передаче по ДСК будет определять не энтропией двоичного алфавита, а эффективной энтропией Hе(A) алфавита или пропускной способностью канала:

|  |  |
| --- | --- |
| Hе(A) = 1 – H(Y|X) | (1.6) |

где H(Y|X) – условная энтропия и

|  |  |
| --- | --- |
| H(Y|X) = –p log2 p – q log2 q. | (1.7) |

1. **Ход работы**

Создать приложение для расчета и анализа параметров и информативных характеристик дискретных ИС, с помощью которого:

а) рассчитать энтропию указанного преподавателем алфавитов: один – на латинице, другой – на кириллице; в качестве входного может быть принят произвольный электронный текстовый документ на основе соответствующего алфавита; частоты появления символов алфавитов оформить в виде гистограмм (можно воспользоваться приложением MSExcel);

С физической точки зрения энтропия алфавита показывает, какое количество информации приходится в среднем на один символ алфавита.

Энтропию алфавита А={ai} по К.Шеннону рассчитывают по следующей формуле:

|  |  |
| --- | --- |
|  | (2.1) |

Функция для расчёта энтропии приведена в листинге 2.1.

public double ShannonEntropy(string s)

{

var map = new Dictionary<char, int>();

foreach (char c in s)

{

if (!map.ContainsKey(c))

map.Add(c, 1);

else

map[c] += 1;

}

double result = 0.0;

int len = s.Length;

var orderkey = from i in map orderby i.Key select i;

foreach (var item in orderkey)

{

using (StreamWriter sw = new StreamWriter("info.log", true, Encoding.Default))

{

sw.Write($"{DateTime.Now} Количество символов {item.Key} = {item.Value} \n");

}

}

foreach (var item in map)

{

var frequency = (double)item.Value / len;

result += frequency \* Math.Log(frequency, 2);

}

return -result;

}

Листинг 2.1 - Функция ShannonEntropy для расчёта энтропии по Шеннону

В результате было получено, что энтропия по Шеннону произвольного текста длиной 1000 знаков на русском языке равна 4,404169644128674, а энтропия по Шеннону произвольного текста на английском языке равна 4,145785814592481.

Для подсчёта частоты появления символов в функции ShannonEntropy также предусмотрена запись в журнал. На основе этих данных были построены диаграммы.

Гистограмма частоты появления символов русского алфавита приведена на рисунке 2.1.

Рисунок 2.1 - Гистограмма частоты появления символов русского алфавита

Гистограмма частоты появления символов латинского алфавита приведена на рисунке 2.2.

Рисунок 2.2 - Гистограмма частоты появления символов латинского алфавита

б) для входных документов, представленных в бинарных кодах, определить энтропию бинарного алфавита;

Для подсчёт энтропии бинарного алфавита использовалась функция ShannonEntropy, представленная в листинге 1.

В результате было получено, что энтропия по Шеннону бинарного алфавита равна 0,9781595325374889.

Гистограмма частоты появления бинарных символов приведена на рисунке 2.3.

Рисунок 2.3. Диаграмма частоты появления бинарных символов

в) используя значения энтропии алфавитов, полученных в пп. а) и б), подсчитать количество информации в сообщении, состоящем из собственных фамилии, имени и отчества (на основе исходного алфавита – а) и в кодах ASCII–б); объяснить полученный результат;

Функция для вычисления количества информации в строке в листинге 2.2.

public double AmountOfInformation(string message, double shannonEntropy)

{

return message.Length \* shannonEntropy;

}

Листинг 2.2 - Функция для вычисления количества информации в строке

В результате было получено, что количество информации в ФИО на латинице равно 78,76993047725713, а в символах ASCII – 215,580862358809. Это можно объяснить тем, что энтропия десятичного алфавита меньше, чем латинского, но длина строки на латинице гораздо меньше длины этой же строки, закодированной символами ASCII так как каждый символ кодируется несколькими цифрами.

г) выполнить задание п. в) при условии, что вероятность ошибочной передачи единичного бита сообщения составляет: 0.1; 0.5; 1.0.

Функция для вычисления количества информации в строке с поправкой на вероятность ошибки представлена в листинге 2.3.

public double AmountOfInformationWithMistake(double entropy, string message, double q)

{

return message.Length\*(entropy - (-(1 - q) \* Math.Log((1 - q), 2) - q \* Math.Log(q, 2)));

}

Листинг 2.3- Функция для вычисления количества информации в строке с поправкой на вероятность ошибки

В результате было получено, что количество информации в ФИО с вероятности ошибки передачи единичного бита 0,1 равно 69,85901419906078, с вероятностью 0,5 - 59,769930477257134, с вероятностью 1 – не будет являться числом поскольку невозможно вычислить логарифм от нуля (а в данном случае вероятность правильной передами равно нулю).

1. **Выводы**

В данной лабораторной работе были изучены теоретические основы теории информации, создано приложение для расчета и анализа параметров и информативных характеристик дискретных ИС. В результате были вычислены энтропии для бинарного, русского и латинского алфавитов, а также количество информации, передаваемой в сообщениях.