«Оценка количества информации в сообщении и эффективное кодирование»

Цель работы

Получение практических навыков численного определения количества информации, содержащегося в сообщении. Освоение методов построения кодов дискретного источника информации используя конструктивный метод, предложенный К. Шенноном и Н. Фано, и метод Хаффмана. На примере показать однозначность раскодирования имеющегося сообщения.

Лабораторная работа состоит из 3 частей, выполняемых последовательно и оцениваемых отдельно.

ЧАСТЬ 1.

Определение количества информации, содержащейся в сообщении

Порядок выполнения части 1 лабораторной работы

1. Создать таблицу (***50 рабочих строк****) в Excel аналогичную рис.1.

Таблица расчета энтропии источника

№ п/п	***Символ***	***Код символа***	***Число вхождений символа в текст***	***Вероятность вхождения символа (p_i)***	***I_i***
1	0				
2	1				
…	…				
50	я				
		Всего символов в тексте (К)			
			Полная вероятность(Р)	(должна получиться «1»)	
				Энтропия источника ($I_{ср}$)	

Рис.1.

****Для заочной формы обучения допускается уменьшить количество строк до 35***

2. Заполнить столбец ***Символ*** следующими значениями:
    - 33 буквы русского алфавита;
    - 10 цифр (0 — 9);
    - Знаки препинания – «.», «,», «:», «;», «-», « », «(».
3. Заполнить столбец ***Код символа*** используя функцию «КОДСИМВ(…)», находящуюся в категории «Текстовые».
4. Выбрать текст на русском языке размером **не менее 10000** символов, указать в отчете данные об использованном тексте (автор, название).
5. Открыв текст и таблицу и используя в Word «***Правка ⇒ Заменить***» заполнить столбец ***Число вхождений символа в текст***. (Предполагается, что других символов в тексте НЕТ.) Сосчитать общее число символов.
6. По формулам заполнить столбцы «p_i» и «I_i ». Сосчитать полную вероятность и энтропию источника.
7. Создать таблицу, аналогичную рис.2 и заполнить ее по формулам.

	Неопределенность	Разрядность кода	Абсолютная избыточность	Относительная избыточность
При кодировании сообщения стандартной кодовой таблицей ASCII				
При использовании равномерного кода, построенного на основе меры Хартли				

Рис.2.

8. Выписать применяемые формулы с расшифровкой используемых символов.

Содержание отчёта

1. Название и цель работы.
2. Название и автор используемого текста
3. Заполненная таблица №1 для 50-ти символов.
4. Заполненная таблица №2.
5. Использованные формулы с определением переменных.
6. Выводы по работе соответственно цели лабораторной работы. Сравнительный анализ таблицы на рис.2.

Приложение к части 1 «Определение количества информации, содержащейся в сообщении»

Основные положения

1. Общие сведения об информации.

Понятие «информация» происходит от латинского слова **informatio**- разъяснение, осведомление, изложение и обозначает одно из основных свойств материи. В рамках науки — информация — первичное, неопределенное понятие. Оно предполагает наличие материального носителя информации, источника информации, передатчика и т.п. Конкретное толкование элементов, связанных с понятием информации, связано с методологией конкретной области науки.

Можно выделить некоторые свойства информации, определяющие смысл этого понятия:

- Информация переносит знания об окружающем мире, которых в рассматриваемой точке не было до получения информации;
- Информация не материальна — она проявляется в форме материальных носителей — дискретных знаков, сигналов или функций времени;
- Информация может быть заключена в знаках или в их взаимном расположении;
- Знаки и сигналы несут информацию только для получателя, который может их распознать.

Термин «информация» имеет много определений. В широком смысле —

Информация— отражение реального мира.

Существует определение термина в узком смысле, применимого к предметной области автоматизированной обработки информации.

Информация — любые сведения, являющиеся объектом хранения, передачи и преобразования.

В процессе передачи информации важно определить следующие понятия:

Сообщение — информация, представленная в определенной форме и предназначенная для передачи. Сообщение представляется последовательностью знаков и сигналов.

Сигнал — процесс, несущий информацию. Таким образом, сигнал служит для переноса информации.

Знак — реально различимые получателем материальные объекты: буквы, цифры, предметы. Знаки служат для хранения информации.

Данные — информация, представленная в формализованном виде и предназначенная для обработки техническими средствами.

Таким образом, любой информационный процесс, может быть представлен как процесс передачи информации от объекта, являющегося источником информации, к получателю. Для обеспечения передачи информации необходим канал связи, некоторая физическая среда, через которую информация, представленная в виде сигналов, передается получателю.

Множество всех знаков и сигналов, использующееся для формирования сообщения, называется алфавит.

Размер (глубина) алфавита ***A*** определяется количеством символов, составляющих алфавит. Если считать, что сообщение передается одним знаком алфавита размером ***A***, всего может быть передано ***N=A*** сообщений.

Из знаков алфавита может быть составлено слово. Если размер слова фиксировано и составляет n знаков, то количество возможных слов ***N*** составленных символов из алфавита ***A***, таким образом, что каждый символ алфавита может входить в слово 0,1,2,…,n , раз определяется

$$N = A^n. \qquad (1)$$

Таким образом, с помощью слов можно представить информацию о любом из ***N*** сообщений.

Выражение (1) позволяет определить размер слова из алфавита **A**, с помощью которого можно представить **N** сообщений

$$n=\lceil \log_A N \rceil. \qquad (2)$$

Мы можем сопоставить тому или иному сообщению комбинацию знаков, тогда при приеме сообщения, зная правила сопоставления, можно распознать сообщение.

Информация всегда представляется в виде сообщения, которое передается некоторой физической средой. Носителем сообщения выступает сигнал, выражающийся в изменении энергии среды передачи информации — канала связи. Для того, чтобы передать информацию по каналу связи необходимо сопоставить исходному сообщению некоторое правило изменения сигнала. Такое правило сопоставления называют кодированием.

Кодирование — представление сообщений в форме, удобной для передачи информации по каналам связи.

Естественно, можно говорить о кодировании на различных этапах передачи информации. Так, например, можно говорить о кодере источника, кодере канала связи и т.д. Принятое сообщение подвергается декодированию.

Декодирование — операция восстановления принятого сообщения. В системе связи необходимо ввести устройства кодирования и декодирования. Очевидно, что правила кодирования и декодирования в системе должны быть согласованы.

Важный вопрос теории передачи и преобразования информации — установление меры, количества и качества информации.

2. Математические меры информации.

Информационные меры, как правило, рассматриваются в двух аспектах синтаксическом и семантическом.

В синтаксическом аспекте сообщения рассматриваются как символы, абстрагированные от содержания и какой-либо ценности. Предметом анализа и оценивания являются частота появления символов, связи между ними, порядок следования, правила построения сообщений. В таком рассмотрении наиболее широко используют ***структурные*** и ***вероятностные*** (статистические) меры.

Структурные меры оценивают строение массивов информации и их измерение простым подсчетом информационных элементов или комбинаторным методом. Структурный подход применяется для оценки возможностей информационных систем вне зависимости от условий их применения.

При статистическом подходе используется понятие энтропии как меры неопределенности, учитывающей вероятность появления и информативность того или иного сообщения. Статистический подход учитывает конкретные условия применения информационных систем.

Семантический подход позволяет выделить полезность или ценность информационного сообщения (в настоящем пособии не рассматривается).

При синтаксическом анализе информация определяется как мера уменьшения неопределенности знаний о каком-либо предмете в познавательном процессе. Если H_1 — исходная (априорная) неопределенность до передачи сообщения, а H_2 — остаточная (апостериорная) неопределенность, характеризующая состояние знания после получения сообщения, то содержащаяся в этом сообщении информация определяется их разностью

$$I=H_1 - H_2. \qquad (3)$$

Известно достаточно большое количество различных мер, различающихся подходом к определению неопределенности в (3). Далее рассматриваются только две из них — структурная аддитивная мера Хартли и вероятностная мера, называемая энтропия, предложенная К.Шенноном.

3. Структурная мера информации. Аддитивная мера Хартли.

Аддитивная мера (мера Хартли) использует понятия глубины A и длины n числа.

Глубина числа — количество символов (элементов), принятых для представления информации. В каждый момент времени реализуется только один какой-либо символ.

Длина n числа — количество позиций, необходимых и достаточных для представления чисел заданной величины.

Эти понятия могут быть распространены и на вариант нечислового сообщения. В этом случае глубина числа тождественна размеру алфавита, а длина числа — разрядности слова при передаче символьного сообщения.

Если сообщение — число, понятие глубины числа будет трансформировано в понятие основания системы счисления. При заданных глубине и длине числа количество чисел, которое можно представить, $N = A^n$. Очевидно, что N однозначно характеризует степень исходной неопределенности. Исходная неопределенность по Хартли определяется

$$H_1 = \log_a N. \quad (4)$$

Неопределенность после получения сообщения, остаточная неопределенность,

$$H_2 = \log_a N^*, \quad (5)$$

где N^* — число возможных значений принятого слова после получения сообщения.

Основание логарифма в (5) определяет только единицы измерения неопределенности. При $a=2$ это двоичная единица информации, называемая бит. При $a = 10$ десятичная (***дит***), при $a = e$ натуральная (***нат***). Далее мы будем всегда пользоваться двоичной единицей.

N^* равно единице, если после получения информации нет неопределенности, т.е. получатель гарантировано получил то сообщение, которое было передано. Если получателю приходится после приема информации выбирать сообщения из некоторого множества, а это происходит тогда, когда в канале связи за счет влияния помех возникают искажения переданного сигнала, то характеризует число возможных сообщений при выборе. Таким образом, если передается символ некоторого алфавита, N^* определяет возможную неоднозначность приема символа за счет искажений в канале связи. В случае измерительного опыта, число N^* — характеризует число возможных значений величины после измерения и определяет погрешность измерения.

Очевидно, что должно быть $N^* < N$, a $N^* = 1$ только в идеальном случае передачи сообщения без потери информации или, что то же самое, измерения некоторой физической величины без ошибок. Количество информации по Хартли оценивается как

$$I=H_1 - H_2 = \log_a N - \log a N^* n = \log_a N/ N^* . \quad (6)$$

Логарифмическая мера, позволяющая, вычислять количество информации, содержащейся в сообщении, переданном числом длиной n и глубиной A:

$$I(q) =\log_2 N=n \log_2 A, \text{бит}. \quad (7)$$

Следовательно, ***1 бит*** информации соответствует одному элементарному событию, которое может произойти или не произойти. Такая мера количества информации удобна тем, что она обеспечивает возможность оперировать мерой как числом. Из сравнения (7) и (2) следует, что численное значение неопределенности определяет число двоичных разрядов, необходимое для кодирования символа алфавита ***A***.

Логарифмическая мера для неопределенности и информации выбрана не случайно. Она оказывается удобной при описании сложных опытов. Допустим, что задача состоит в одновременном приеме информации от двух источников, не зависящих друг от друга. При этом N_1 и n_1 — число возможных сообщений до и после приема информации от первого источника, а — N_2 и n_2 от второго. Пусть H_{11} и H_{12} — исходная неопределенность знания первого и второго сообщения, соответственно, первого и второго источника. Естественно потребовать, чтобы общая неопределенность знания о двух сообщениях определялась суммой неопределенностей каждого, т.е. мера должна обладать свойством аддитивности

$$H = H_{11} + H_{12}.$$

Число возможных сочетаний двух независимых величин из множеств $N_1 N_2$ $N = N_1 N_2$.

Тогда исходная неопределенность $H = H_{11} + H_{12}$, , аналогично остаточная неопределенность $H = H_{21} + H_{22}$.

При наличии нескольких источников информации общее количество информации

$$I(q_1, q_2, \ldots, q_n) = I(q_1) + I(q_2) + \ldots + I(q_k), \quad (8)$$

где $I(q_k)$ — количество информации от источника k.

Логарифмическая мера информации позволяет измерять количество информации и широко используется на практике. Однако всегда надо учитывать, что все сообщения в этой мере полагаются равновероятными и независимыми. Эти допущения приводит на практике к существенно завышенным оценкам.

Примечание. *Для рассмотрения дальнейшего материала необходимо использовать понятие* ***«вероятность события»****. Под вероятностью события (см., например, Лютикас В.С. Факультативный курс по математике. Теория вероятностей. М.: Просвещение, 1990.) принимается постоянная величина, около которой группируются значения частоты появление некоторого события, например, передачи одного из символов алфавита. Если частота появления любого символа алфавита при передаче длинной последовательности символов одинакова, то говорят о равновероятных событиях, символах, сообщениях и т.п. Независимыми сообщения*

полагают, если вероятности их передачи не зависят от того, какие сообщения были переданы ранее.

4. Статистическая мера информации.

В статистической теории информации вводится более общая мера количества информации, в соответствии с которой рассматривается не само событие, а информация о нем. Этот вопрос глубоко проработан К. Шенноном в работе «Избранные труды по теории информации». Если появляется сообщение о часто встречающемся событии, вероятность появления которого близка к единице, то такое сообщение для получателя малоинформативное. Столь же мало информативны сообщения о событиях, вероятность появления которых близка к нулю.

События можно рассматривать как возможные исходы некоторого опыта, причем все исходы этого опыта составляют ансамбль, или полную группу событий. К. Шеннон ввел понятие неопределенности ситуации, возникающей в процессе опыта, назвав ее энтропией. Энтропия ансамбля есть количественная мера его неопределенности и, следовательно, информативности, количественно выражаемая как средняя функция множества вероятностей каждого из возможных исходов опыта.

Поясним содержание статистической меры на следующем частном случае. Пусть выполняется посимвольная передаче текста, состоящего из символов алфавита A. Текст составлен из K символов алфавита. Опыт состоит в передаче очередного символа текста. Так как в один момент времени может быть передан любой символ алфавита, всего возможно A исходов опыта. Очевидно, что одни символы в тексте будут появляться часто, а другие — реже. Различные символы несут разную информацию. Обозначим через k_i количество появления символа в тексте, а количество вносимой этим символом информации как I_i. Будем полагать, что передаваемые символы независимы, т.е. передача i –того символа происходит с вероятностью, независящей от того, какой символ был передан ранее. Это означает, информация, вносимая символом постоянна для любых сочетаний символов. Тогда средняя информация, доставляемая одним опытом,

$$I_{cp} = (k_1 I_1 + k_2 I_2 + \ldots + k_A I_A)/K. \quad (9)$$

Но количество информации в каждом исходе связано с его вероятностью p_i , и выражается в двоичных единицах (битах) как

$$I_i = \log_2 (1/p_i) = -\log_2 p_i .$$

Тогда

$$I_{cp} = [k_1 (-\log_2 p_1) + \ldots + k_A (-\log_2 p_A)]/K. \quad (10)$$

Выражение (10) можно записать также в виде

$$I_{cp} = k_1/K (-\log_2 p_1) + \ldots + k_A/K (-\log_2 p_A). \quad (11)$$

Но отношения ***n/K*** представляют собой частоты повторения исходов, а, следовательно, могут быть заменены их вероятностями:

$$p_i = k_i/K ,$$

Тогда средняя информация в битах

$$I_{cp} = p_1 (-\log_2 p_1) + \ldots + p_A (-\log_2 p_A),$$

или

$$I_{cp} = \sum p_i (-\log_2 p_i) = H \qquad (12)$$

Полученную величину Н называют энтропией. Энтропия обладает следующими свойствами:

1. Энтропия всегда неотрицательна, так как значения вероятностей выражаются величинами, не превосходящими единицу, а их логарифмы — отрицательными числами или нулем, так что члены суммы (12) — неотрицательны.

2. Энтропия равна нулю в том крайнем случае, когда одно из pi , равно единице, а все остальные — нулю. Это тот случай, когда об опыте или величине все известно заранее и результат не дает новую информацию.

3. Энтропия имеет наибольшее значение, когда все вероятности равны между собой:

$$p_1 = p_2 = \ldots = p_i = 1/A.$$

Приэтом$H = -\log_2(1/A) = \log_2 A = H_{max}$.

4. Энтропия объекта BC, состояния которого образуются совместной реализацией состояний B и C, равна сумме энтропии исходных объектов ***B*** и ***C***, т. е. $H(BC) = H(B) + H(C)$.

Если все события равновероятны и статистически независимы, то оценки количества информации, по Хартли и Шеннону, совпадают. Это свидетельствует о полном использовании информационной емкости системы. В случае неравных вероятностей количество информации, по Шеннону, меньше информационной емкости системы. Максимальное значение энтропии достигается при $p=0,5$, когда два состояния равновероятны. При вероятностях $p=0$ или $p=1$, что соответствует полной невозможности или полной достоверности события, энтропия равна нулю.

Наибольшее количество информации получается тогда, когда полностью снимается неопределенность, причем эта неопределенность была наибольшей — вероятности всех событий были одинаковы. Это соответствует максимально возможному количеству информации, оцениваемому мерой Хартли:

$$I_x = \log_2 N = \log_2 (1/p) = -\log_2 p = H_{max},$$

где N — число событий; p — вероятность их реализации в условиях равной вероятности событий, H_{max}— максимальное значение неопределенности, равное энтропии равновероятностных событий.

Абсолютная избыточность информации $D_{авс}$ представляет собой разность между максимально возможным количеством информации и энтропией:

$$D_{авс} = I_x - H, \text{ или } D_{авс} = H_{max} - H . \quad (13)$$

Пользуются также понятием относительной избыточности

$$D = (H_{max} - H)/H_{max}. \quad (14)$$

Рассмотренные информационные меры в полной мере применимы для оценки количества информации при передаче и хранении информации в вычислительных системах и цифровых системах связи. Если информация передается с использованием некоторого алфавита А то передачу каждого символа можно рассматривать как опыт, имеющий А возможных исходов. В длинном сообщении, например, при передаче текста размером К символов, различные символы алфавита могут появляться различное число раз. Мы можем говорить о частоте появления символов в сообщении, которая с увеличением К стремится к вероятности появления конкретного символа в сообщении.

Информационные меры имеют важное значение при определении характеристик памяти ЭВМ, пропускной способности каналов связи и во многих других приложениях информатики.

Часть 2.

«Кодирование дискретных источников информации методом Шеннона-Фано»

Порядок выполнения части 2 лабораторной работы

Исходными данными для данной лабораторной работы являются результаты статистической обработки текста, выполненной в предыдущей лабораторной работе. Из лабораторной работы «Определение количества информации, содержащегося в сообщении» для данной работы необходимо взять:

1) список символов данного текста;
2) оценку вероятностей появления символов в тексте;
3) значение энтропии источника.

Расчеты рекомендуется выполнять в табличной форме, используя MSExcel.

1. Отсортировать символы в порядке убывания их вероятности появления в тексте.
2. Построить один из возможных вариантов по правилу Шеннона-Фано для посимвольного кодирования заданного текста.
3. Определить энтропию и среднее количество двоичных разрядов, необходимых для передачи текста при использовании эффективных кодов.
4. Проверить возможность однозначного декодирования полученных кодов, рассмотрев пример передачи слова, состоящего из ***не менее*** 10 символов.

Содержание отчёта

1. Название и цель работы.
2. Заполненная таблица для 50-ти символов, содержащая:
   - список символов;
   - значения вероятностей;
   - кодовые комбинации;
   - ступени деления.
3. Значение средней информации в битах.
4. Описание применяемых формул.
5. Составленное сообщение, содержащее не менее 10 символов и кодовая строка.
6. Описание декодирования данного сообщения любым способом.
7. Таблица, содержащая
   - список символов;

- значения вероятностей;
- 49 шагов суммирования вероятностей.

8. Значение средней информации в битах, вычисленное по составленной таблице кодов.
9. Описание применяемых формул.
10. Рисунок кодового дерева, с полученными значениями и подписанными символами.
11. Таблица полученных кодов.
12. Составленное сообщение, содержащее не менее 10 символов и кодовая строка.
13. Описание декодирования данного сообщения.
14. Выводы по работе соответственно цели лабораторной работы.

Приложение к части 2 лабораторной работы

Кодирование дискретных источников информации методом Шеннона-Фано

Основные положения

При кодировании дискретных источников информации часто решается задача уменьшения избыточности, т.е. уменьшения количества символов, используемых для передачи сообщения по каналу связи. Это позволяет повысить скорость передачи за счет уменьшения количества передаваемой информации, а точнее, за счет отсутствия необходимости передачи избыточной информации. В системах хранения уменьшение избыточности позволяет снизить требования к информационной емкости используемой памяти.

Для передачи и хранения информации, как правило, используется двоичное кодирование. Любое сообщение передается в виде различных комбинаций двух элементарных сигналов. Эти сигналы удобно обозначать символами 0 и 1. Тогда кодовое слово будет состоять из последовательностей нулей и единиц.

Если алфавит A состоит из N символов, то для их двоичного кодирования необходимо слово разрядностью n, которая определяется

$$n = \lceil \log_2 N \rceil.$$

Это справедливо при использовании стандартных кодовых таблиц, например, ASCII, KOI-8 и т.п., обеспечивающих кодирование до 256 символов.

Если в используемом алфавите символов меньше, чем используется в стандартной кодовой таблице, то возможно использование некоторой другой таблицы кодирования, позволяющей уменьшить количество двоичных разрядов, используемых для кодирования любого символа. Это, в определенном смысле, обеспечивает сжатие информации.

Например, если необходимо передавать или хранить сообщение, состоящее из символов кириллицы, цифр и семи символов разделителей {«.», «,», «:», «;», «!», « кавычки », «?»} (всего 50 символов), мы можем воспользоваться способами кодирования:

- Кодировать каждый символ в соответствии со стандартной кодовой таблицей, например, KOI-8R. По этой таблице каждый символ будет представляться 8 битовым (байт) кодовым словом, $n_1 = 8$;
- Составить и использовать отдельную кодовую таблицу, это может быть некоторый усеченный вариант стандартной таблицы, не учитывающую возможность кодирования символов, не входящих в передаваемое сообщение, тогда необходимый размер кодового слова

$$n_2 = \lceil \log_2 N \rceil = \lceil \log_2 50 \rceil = 6.$$

Очевидно, передача сообщения с помощью такого кода будет более эффективной, т.к. будет требовать меньшего количества бинарных сигналов при кодировании. Можно говорить о том, что при таком кодировании мы не передаем избыточную информацию, которая была бы в восьмибитовом кодировании;

- Использовать специальный код со словом переменной длины, в котором символы, появляющиеся в сообщении с наибольшей вероятностью, будут закодированы короткими словами, а наименее вероятным символам сопоставлять длинные кодовые комбинации. Такое кодирование называется эффективным.

Эффективное кодирование базируется на ***основной теореме Шеннона*** для каналов без шума, в которой доказано, что ***сообщения, составленные из букв некоторого алфавита, можно закодировать так, что среднее число двоичных символов на букву будет сколь угодно близко к энтропии источника этих сообщений, но не меньше этой величины.***

Теорема не указывает конкретного способа кодирования, но из нее следует, что при выборе каждого символа кодовой комбинации необходимо стараться, чтобы он нес максимальную информацию. Следовательно, каждый элементарный сигнал должен принимать значения 0 и 1 по возможности с равными вероятностями и каждый выбор должен быть независим от значений предыдущих символов.

При отсутствии статистической взаимосвязи между кодируемыми символами конструктивные методы построения эффективных кодов были даны впервые К.Шенноном и Н.Фано. Их методики существенно не различаются, поэтому соответствующий код получил название кода Шеннона-Фано.

Код строится следующим образом:

буквы алфавита сообщений выписываются в таблицу в порядке убывания вероятностей. Затем они разделяются на две группы так, чтобы суммы вероятностей в

каждой из групп были по возможности одинаковы. Всем буквам верхней половины в качестве первого символа приписывается 1, а всем нижним — 0. Каждая из полученных групп, в свою очередь, разбивается на две подгруппы с одинаковыми суммарными вероятностями и т. д. Процесс повторяется до тех пор, пока в каждой подгруппе останется по одной букве.

Рассмотрим алфавит из восьми букв (табл. 2.1). Ясно, что при обычном (не учитывающем статистических характеристик) кодировании для представления каждой буквы требуется $n_2 = 3$ символа. В табл.2.1 приведен один из возможных вариантов кодирования по сформулированному выше правилу.

Таблица 2.1

Символы	Вероятности $p(a_i)$	Кодовые комбинации	1 ступень	2 ступень	3 ступень	4 ступень	5 ступень
a_1	0.22	11					
a_2	0.20	10					
a_3	0.16	011					
a_4	0.16	010					
a_5	0.10	001					
a_6	0.10	0001					
a_7	0.04	00001					
a_8	0.02	00000					

Очевидно, для указанных вероятностей можно выбрать другое разбиение на подмножества не нарушая алгоритма Шеннона-Фано. Такой пример приведен в табл.2.2.

Таблица 2.2

Символы	Вероятности $p(a_i)$	Кодовые комбинации	1 ступень	2 ступень	3 ступень	4 ступень	5 ступень
a_1	0.22	11					
a_2	0.20	101					
a_3	0.16	100					
a_4	0.16	01					
a_5	0.10	001					
a_6	0.10	0001					
a_7	0.04	00001					
a_8	0.02	00000					

Сравнивая приведенные таблицы, обратим внимание на то, что по эффективности полученные коды различны. Действительно, в табл.2.2 менее вероятный символ a_4 будет

закодирован двухразрядным двоичным числом, в то время как a_2 , вероятность появления которого в сообщении выше, кодируется трехразрядным числом.

Таким образом, рассмотренный алгоритм Шеннона-Фано не всегда приводит к однозначному построению кода. Ведь при разбиении на подгруппы можно сделать большей по вероятности как верхнюю, так и нижнюю подгруппу.

Энтропия набора символов в рассматриваемом случае определяется как

$$I_{cp} = -\sum_{1}^{8} p(a_i)\log_2 p(a_i) \approx 2{,}76.$$

Напомним, что смысл энтропии в данном случае, как следует из теоремы Шеннона, — наименьшее возможное среднее количество двоичных разрядов, необходимых для кодирования символов алфавита размера восемь с известными вероятностями появления символов в сообщении.

Мы можем вычислить среднее количество двоичных разрядов, используемых при кодировании символов по алгоритму Шеннона-Фано

$$n_{cp} = \sum_{1}^{A} p(a_i)n(a_i),$$

где: A — размер (или объем) алфавита, используемого для передачи сообщения;

$n(a_i)$ — число двоичных разрядов в кодовой комбинации, соответствующей символу a_i.

Таким образом, мы получим для табл.1 $n_{cp} = 2{,}84$, а для табл.2 $n_{cp} = 2{,}80$. Построенный код весьма близок к наилучшему эффективному коду по Шеннону, но не является оптимальным. Должен существовать некоторый алгоритм позволяющий выполнить большее сжатие сообщения.

Пример декодирования сообщения

Рассмотрим пример сообщения, созданного из имеющихся символов. Пусть передано сообщение $a_1, a_5, a_3, a_7, a_2, a_3$.

При кодировании, используя табл.1 получим следующую последовательность:

1100101100001101011

Получив сообщение подобного вида, необходимо её декодировать, чтобы прочитать сообщение. Считаем, что получатель имеет таблицу кодировки символов, идентичную с отправителем.

Возможный способ декодирования представлен на таблице 2.3:

Таблица 2.3

шаг	***комбинация***	***кол-во символов***	***символ***
1	11	1	a_1
2	00	4	-
3	001	1	a_5
4	01	2	-
5	011	1	a_3
…			

Декодирование производиться с наименьшей длины кодового слова — в нашем случае — 2, — получается таблица (см.выше), с итоговым результатом, аналогичным закодированному:

Часть 3. «Кодирование дискретных источников информации методом Д.Хаффмана»

Порядок выполнения части 3 лабораторной работы

Из лабораторной работы «Кодирование дискретных источников информации методом Шеннона-Фано» необходимо взять вычисленное значение средней информации.

Расчеты рекомендуется выполнять в табличной форме, используя MSExcel.

1. Отсортировать символы в порядке убывания их вероятности появления в тексте.
2. Построить таблицу по правилу Д. Хаффмана для посимвольного кодирования заданного текста (См. Табл.3.1).
3. Определить энтропию и среднее количество двоичных разрядов, необходимых для передачи текста при использовании эффективных кодов.
4. Построить кодовое дерево (См.рис.3.1).
5. Создать таблицу кодов.
6. Проверить возможность однозначного декодирования полученных кодов, рассмотрев пример передачи слова, состоящего из не менее 10 символов.

Содержание отчёта

1. Название и цель работы.
2. Заполненная таблица для 50-ти символов, содержащая:
   - список символов;
   - значения вероятностей;
   - кодовые комбинации;
   - ступени деления.
3. Значение средней информации в битах.
4. Описание применяемых формул.
5. Составленное сообщение, содержащее не менее 10 символов и кодовая строка.
6. Описание декодирования данного сообщения любым способом.
7. Таблица, содержащая

- список символов;
- значения вероятностей;
- 49 шагов суммирования вероятностей.

8. Значение средней информации в битах, вычисленное по составленной таблице кодов.
9. Описание применяемых формул.
10. Рисунок кодового дерева, с полученными значениями и подписанными символами.
11. Таблица полученных кодов.
12. Составленное сообщение, содержащее не менее 10 символов и кодовая строка.
13. Описание декодирования данного сообщения.
14. Выводы по работе соответственно цели лабораторной работы.

Кодирование дискретных источников информации по методики Д.Хаффмана

Основные положения

От недостатка неоднозначного кодирования, рассмотренного в предыдущей лабораторной работе алгоритма свободна методика Д.Хаффмана. Она гарантирует однозначное построение кода с наименьшим для данного распределения вероятностей средним числом двоичных разрядов на символ.

Для двоичного кода алгоритм Хаффмана сводится к следующему:

Шаг 1. Символы алфавита, составляющего сообщение, выписываются в основной столбец в порядке убывания вероятностей. Два последних символа объединяются в один вспомогательный, которому приписывается суммарная вероятность.

Таблица 3.4

Символы	**Вероятности $p(a_1)$**	**Вспомогательные вычисления**													
			Шаг 1		**Шаг 2**		**Шаг 3**		**Шаг 4**		**Шаг 5**		**Шаг 6**		**Шаг 7**
a_1	0,22		0,22		0,22		0,26		0,32		0,42		0,58		1,0
a_2	0,20		0,20		0,20		0,22		0,26		0,32		0,42		
a_3	0,16		0,16		0,16		0,20		0,22		0,26				
a_4	0,16		0,16		0,16		0,16		0,20						
a_5	0,10		0,10		0,16		0,16								
a_6	0,10		0,10		0,10										
a_7	0,04		0,06												
a_8	0.02														

Шаг 2. Вероятности символов, не участвовавших в объединении, и полученная суммарная вероятность снова располагаются в порядке убывания вероятностей в дополнительном столбце, а

две последних объединяются. Процесс продолжается до тех пор, пока не получим единственный вспомогательный символ с вероятностью, равной единице.

Эти два шага алгоритма иллюстрирует Табл.3.4 для уже рассмотренного случая кодирования восьми символов.

Шаг 3. Строится кодовое дерево и, в соответствии с ним, формируются кодовые слова, соответствующие кодируемым символам.

Поясним принципы выполнения последнего шага алгоритма. Для составления кодовой комбинации, соответствующей данному сообщению, необходимо проследить путь перехода сообщений по строкам и столбцам таблицы. Для наглядности строится кодовое дерево (рис.3.1). Из точки, соответствующей вероятности 1, направляются две ветви. Ветви с большей вероятностью присваивается символ 1, а с меньшей — символ 0. Такое последовательное ветвление продолжаем до тех пор, пока не дойдем до каждого символа.

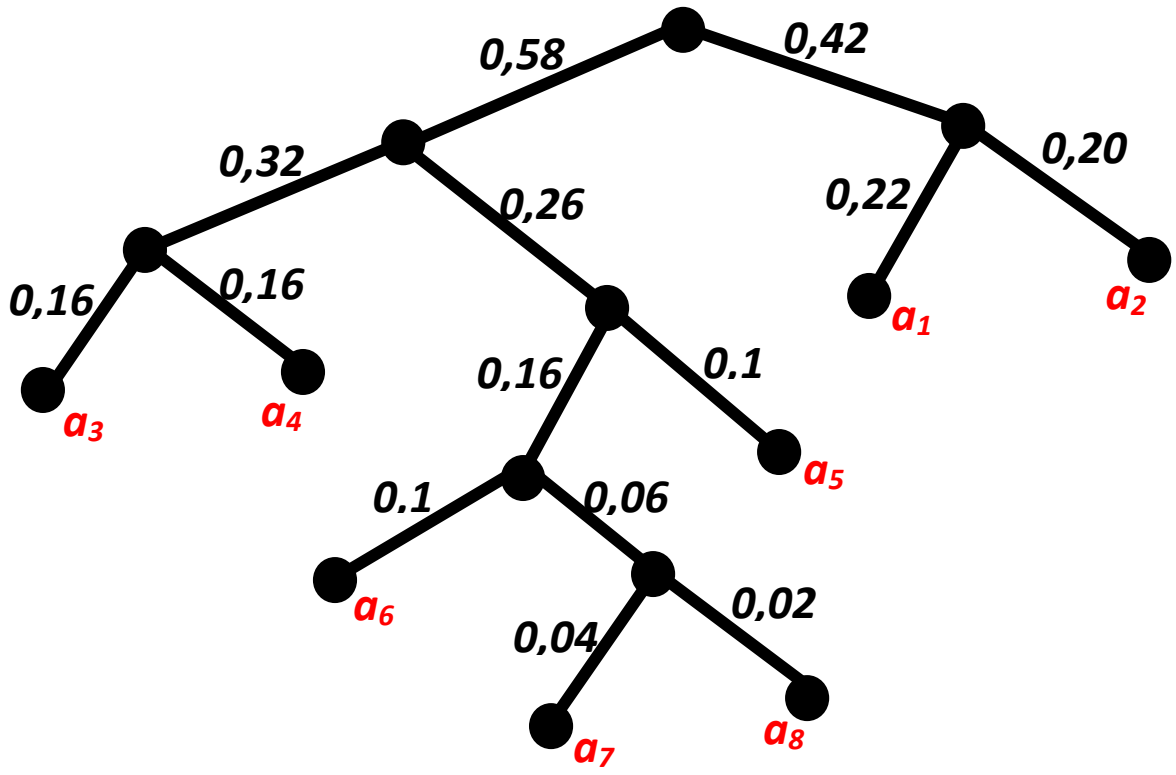

Рис.3.1 *Кодовое дерево*

Таким образом, символам источника сопоставляются концевые вершины дерева. Кодовые слова, соответствующие символам, определяются путями из начального узла дерева к концевым. Каждому ответвлению влево соответствует символ 1 в результирующем коде, а вправо — символ 0.

Поскольку только концевым вершинам кодового дерева сопоставляются кодовые слова, то ни одно кодовое слово не будет началом другого. Тем самым гарантируется возможность разбиения последовательности кодовых слов на отдельные кодовые слова.

Теперь, двигаясь по кодовому дереву сверху вниз, можно записать для каждой буквы соответствующую ей кодовую комбинацию (см.табл.3.5):

Таблица 3.5

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
01	00	111	110	100	1011	10101	10100