Министерство науки и высшего образования
Российской Федерации
Санкт-Петербургский государственный университет
аэрокосмического приборостроения

А.А.Ключарёв

Основы теории информации

Оценка количества информации в сообщении и эффективное кодирование

Методические указания по выполнению контрольной работы для
студентов заочной формы обучения направления 09.03.04
"Программная инженерия"

Санкт-Петербург
2022

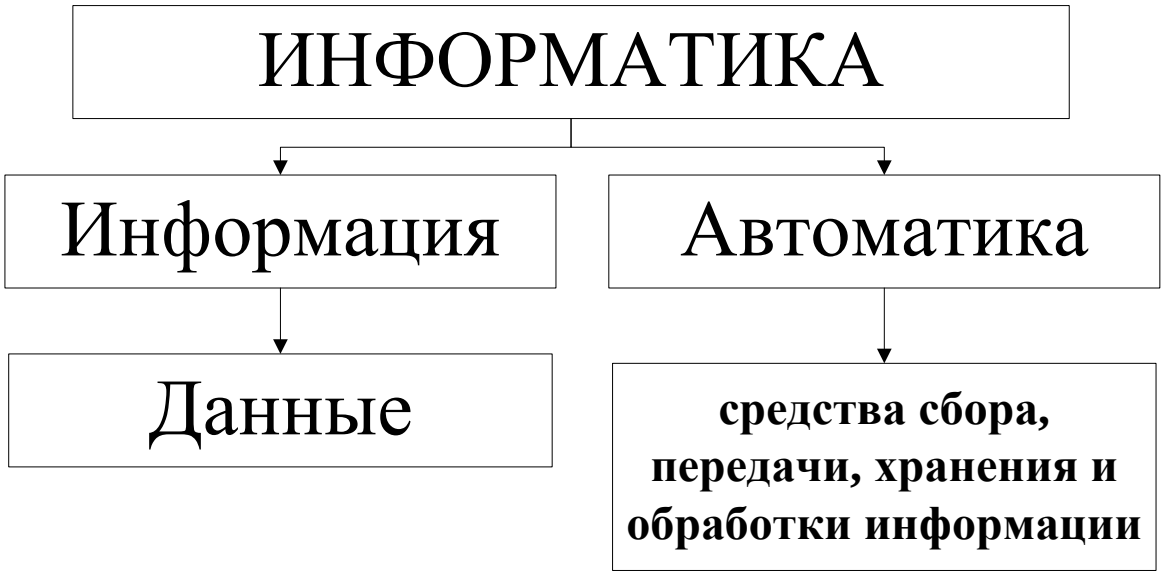
1. ОСНОВНЫЕ ТЕОРЕТИЧЕСКИЕ ПОЛОЖЕНИЯ

1.1. Общие сведения об информации.

Информа́тика (фр. Informatique; англ.Iinformatics или Computer science) — наука о методах и процессах сбора, хранения, обработки, передачи, анализа и оценки информации с применением компьютерных технологий, обеспечивающих возможность её использования для принятия решений (Большая российская энциклопедия, 2008.).

Слово "информатика" образовано из двух слов "информация" и "автоматика".

В англоязычных странах, особенно в США, вместо термина "informatics" обычно используют "computer science", то есть компьютерная наука.



По "автоматикой" в данном контексте подразумеваются средства сбора, передачи, хранения и обработки информации, которые, в свою очередь, можно разделить на технические средства (Hardware) и программное обеспечение, реализующее заданные алгоритмы обработки информации (Software). Такое разделение обусловлено тем, что основным средством, решающим задачи информатики являются вычислительные устройства и системы.

Понятие «информация» происходит от латинского слова informatio- разъяснение, осведомление, изложение и обозначает одно из основных свойств материи. Термин "информация" является первичным и не имеет конкретного определения. Его смысл различен в зависимости от конкретной предметной области, в которой он используется и решаемыми задачами. Например, слово "информация" по различному воспринимается в журналистике, медицине, технике и в других областях, поэтому в литературе можно встретить множество определений этого понятия.

Применительно к задачам, решаемым в области информационных технологий будем полагать, что

Информация — любые сведения об объекте окружающего мира, являющиеся объектом хранения, передачи, преобразования и обработки.

Можно выделить некоторые свойства информации, определяющие смысл этого понятия:

- Информация переносит знания об окружающем мире, которых в рассматриваемой точке не было до получения информации. Если у получателя информации уже есть определенные сведения об объекте, то для него эти сведения уже не являются информацией;
- Информация не материальна и, следовательно, непосредственно ее нельзя передать - переместить от источника к получателю информации;

- Информация может быть заключена в символах (знаках) или их последовательности в определенном порядке - взаимном расположении. Эта последовательность называется ***сообщение***;
- Сообщение может быть передано по каналу связи только с помощью материального носителя. Таким является ***сигнал*** - изменение энергии, однозначно связанное с передаваемыми символами, составляющими сообщение;
- Символы и сигналы несут информацию только для получателя, который может их распознать.

В процессе передачи информации важно определить следующие понятия:

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, использующееся для формирования сообщения, называется алфавит.

Размер (глубина) алфавита A определяется количеством символов, составляющих алфавит. Знаками алфавитом размером A может быть передано N 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

$$N = A^n. \qquad (1.1)$$

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

Выражение (1.1) позволяет определить размер слова из алфавита A, с помощью которого можно представить N сообщений

$$n = \lceil \log_A N \rceil. \qquad (1.2)$$

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.

1.2. 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

$$\mathbf{I=H_1 - H_2.} \qquad (1.3)$$

Известно достаточно большое количество различных мер, различающихся подходом к определению неопределенности в (1.3). Далее рассматриваются только две из них – структурная аддитивная мера Хартли и вероятностная мера, называемая энтропия, предложенная К.Шенноном.

1.3. **Структурная мера информации. Аддитивная мера Хартли.**

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

Глубина **q**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, $\mathbf{N = q^n}$.

Величина N не удобнадля оценки информационной емкости. Исходная неопределенность по Хартли характеризуется логарифмической функцией

$$\mathbf{H_1 = \log_a N}. \qquad (1.4)$$

Неопределенность после получения сообщения, остаточная неопределенность,

$$\mathbf{H_2 = \log_a n}, \qquad (1.5)$$

где **n**– число возможных значений принятого символа после получения сообщения. В случае измерительного опыта, число **n**-характеризует число возможных значений величины после измерения и определяет погрешность измерения. Если передается символ некоторого алфавита, **n** определяет возможную неоднозначность приема символа за счет искажений в канале связи.

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

$$\mathbf{I=H_1 - H_2 = \log_a N - \log_a n = \log_a N/n.} \quad (1.6)$$

Основание логарифма в 1.5 определяет только единицы измерения неопределенности. При **a=2**это двоичная единица информации, называемая бит. При **a = 10**десятичная (дит), при **a =e**натуральная (нит). Далее мы будем всегда пользоваться двоичной единицей.

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

$$\mathbf{I(q),=\log_2 N=n \log_2 q} \text{ , бит.} \quad (1.7)$$

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. Количество информации при этом эквивалентно количеству двоичных символов 0 или 1.

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 $\mathbf{N_1}$ и $\mathbf{n_1}$ – число возможных сообщение до и после приема информации от первого источника, а - $\mathbf{N_2}$ и $\mathbf{n_2}$от второго. Пусть $\mathbf{H_{11}}$и$\mathbf{H_{12}}$ – 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**

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

Число возможных сочетаний двух независимых величин из множеств $\mathbf{N_1 N_2}$

$$\mathbf{N= N_1 N_2.}$$

Тогда исходная неопределенность $\mathbf{H = H_{11} + H_{12},}$ **,** аналогично остаточная неопределенность$\mathbf{H = H_{21} + H_{22}.}$При наличии нескольких источников информации общее количество информации

$$\mathbf{I(q_1, q_2, ...,q_n)= I(q_1)+ I(q_2)+...+I(q_n),} \quad (1.8)$$

где $\mathbf{I(q_k)}$ — 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.*

1.4. 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 жемалоинформативны сообщения о событиях, вероятность появления которых близка к нулю.

События можно рассматривать как возможные исходы некоторогоопыта, 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.

Пусть выполняется **K**опытов. В каждом опыте возможен один из **k**различных исходов. Некоторый **i**-й исход повторяется $\mathbf{n_i}$ раз и вносит информацию, количество которой оценивается как $\mathbf{I_i}$. Тогда средняя информация, доставляемая одним опытом,

$$\mathbf{I_{cp}=(n_1I_1+n_2I_2+\ldots+n_kI_k)/K.} \qquad (1.9)$$

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

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

Тогда

$$\mathbf{I_{cp}=[n_1 (-\log_2 p_1)+\ldots+n_k (-\log_2 p_k)]/K.} \qquad (1.10)$$

Выражение (1.10) можно записать также в виде

$$\mathbf{I_{cp} =n_1/K (-\log_2 p_1)+\ldots+n_k/K (-\log_2 p_k).} \qquad (1.11)$$

Но отношения **n/K** представляют собой частоты повторения исходов, а, следовательно, могут быть заменены их вероятностями: $\mathbf{n_i/K=p_i}$, поэтому средняя информация в битах

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

или

$$\mathbf{I_{cp} =\sum p_i (-\log_2 p_i) = H} \qquad (1.12)$$

Полученную величину **H** 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 (1.12) — неотрицательны.

2. Энтропия равна нулю в том крайнем случае, когда одно из $\mathbf{p_i}$, 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:

$\mathbf{p_1 =.p_2=\ldots =p_i =1/k}$. При этом

$$\mathbf{H=-\log_2(1/k)=\log_2 k}.$$

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

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.

Количество информации только тогда равно энтропии, когда неопределенность ситуации снимается полностью. В общем случае нужно считать, что количество информации есть уменьшение энтропии вследствие опыта или какого-либо другого акта познания. Если неопределенность снимается полностью, то информация равна энтропии –

$$\mathbf{I = H}\ .$$

В случае неполного разрешения имеет место частичная информация, являющаяся разностью между начальной и конечной энтропией: $\mathbf{I = H_1 + H_2}$.

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:

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

где **N**— число событий; **p** — вероятность их реализации в условиях равной вероятности событий.

Таким образом, $\mathbf{I_x = H_{max}}$.

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

$$\mathbf{D_{авс} = I_x - H}\ , \text{ или } \mathbf{D_{авс} = H_{max} - H}\ . \qquad (1.13)$$

Пользуются также понятием относительной избыточности

$$\mathbf{D = (H_{max} - H)/H_{max}}. \qquad (1.14)$$

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 **A** то передачу каждого символа можно рассматривать как опыт, имеющий **A** возможных исходов. В длинном сообщении, например, при передаче текста размером **K** символов, различные символы алфавита могут появляться различное число раз. Мы можем говорить о частоте появления символов в сообщении, которая с увеличением **K**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.

1.5. Уменьшение избыточности при передаче и хранении информации. эффективное кодирование

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

$\mathbf{n = \lceil \log_2 N \rceil}$.

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, $\mathbf{n_1 = 8}$;
- 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

  $\mathbf{n_2 = \lceil \log_2 N \rceil == \lceil \log_2 50 \rceil = 6.}$

  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.

Из теоремы Шеннона следует физический смысл любой информационной меры:

Количество информации соответствует средней длине слова из символов данного алфавита необходимой для кодирования любого сообщения из данного набора.

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.1.1). Ясно, что при обычном (не учитывающем статистических характеристик) кодировании для представления каждой буквы требуется $\mathbf{n_2 = 3}$ символа. В табл.1.1 приведен один из возможных вариантов кодирования по сформулированному выше правилу.

Таблица 1.1

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

Очевидно, для указанных вероятностей можно выбрать другое разбиение на подмножества не нарушая алгоритма Шеннона-Фано. Такой пример приведен в табл.1.2.

Сравнивая приведенные таблицы, обратим внимание на то, что по эффективности полученные коды различны. Действительно, в табл.1.2 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.

Таблица 1.2

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

Энтропия набора символов в рассматриваемом случае определяется как

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

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

$$I_{cp} = -\sum_{1}^{A} p(a_i)n(a_i), \qquad (1.15)$$

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 $\boldsymbol{I_{cp}}$=2,84, а для табл.2 $\boldsymbol{I_{cp}}$=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.

От недостатка рассмотренного алгоритма свободна методика Д. Хаффмена. 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.

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 единственную вспомогательный символ с вероятностью, равной единице.

Эти два шага алгоритма иллюстрирует табл.1.3 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.1.1). 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.

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.

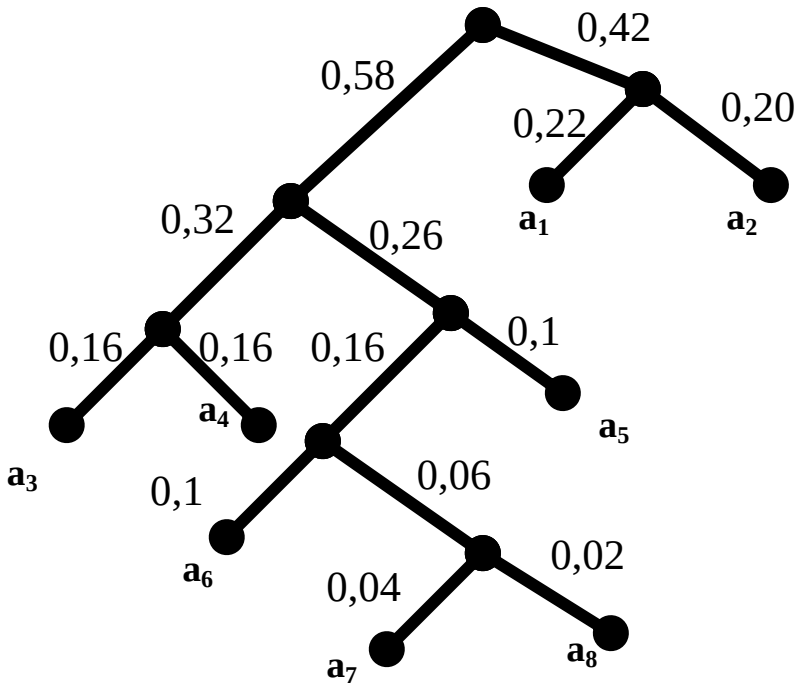


Рис.1.1. Кодовое дерево

Теперь, двигаясь по кодовому дереву сверху вниз, можно записать для каждой буквы соответствующую ей кодовую комбинацию:

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

Покажем, что использованный алгоритм позволяет однозначно декодировать полученное сообщение.

Пусть передаваемое сообщение a_1,a_5,a_3,a_7,a_8. В результате кодирования по алгоритму Хаффмана получим следующую кодовую последовательность:

011001111010110100.

При приеме неизвестно, каким образом эту последовательность надо разбить на кодовые слова и, соответственно, получить последовательность переданных символов.

Просматриваем принятую последовательность слева направо, учитывая, что наибольшая длина кодового слова равна 5. Из первых пяти принятых элементов следует, что комбинация 01100 не встречается ни в одном кодовом слове. Единственное правильное решение, что первым был передан символ a_1. Аналогично, продолжая просмотр с третьего слева элемента, определяем второй символ - a_5. Продолжая просмотр декодируем все сообщение - **.**a_1,a_5,a_3,a_7,a_8**.**

Таким образом, в передаваемой последовательности нет необходимости указывать разделители между отдельными кодовыми словами.

Таблица 1.3

Кодирование по методу Хаффмана

Символы	Вероятности $p(a_i)$	Вспомогательные столбцы						
		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
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	0,1	0,16	0,16				
a_6	0.10	0,1	0,10					
a_7	0.04	0,6						
a_8	0.02							

2. КОНТРОЛЬНАЯ РАБОТА

Целью выполнения контрольной работы является получение практических навыков численного определения количества информации, содержащегося в сообщении и освоение приемов эффективного кодирования. В качестве исходного материала студент должен выбрать произвольный текст на русском языке, желательно без рисунков, таблиц и формул.

2.1. **Определение количества информации, содержащегося в сообщении**

2.1.1. На листе книги MS Excel составьте следующую таблицу.

Таблица 2.1

Результаты работы

№ п/п	символ	ASCII код символа	Число вхождений символа в текст	p	H_i
1					
…	…				
255	я				
Всего символов в тексте					
Неопределенность при использовании стандартной кодовой таблицы					
Неопределенность по Хартли				Энтропия источника	

Для составления перечня символов кодовой таблицы рекомендуется воспользоваться функцией СИМВОЛ(). Исключите из таблицы строки, соответствующие управляющим символам и прописным символам латинского алфавита и кириллицы.

2.1.2. Для выбранного студентом текста (не менее 1000 символов) заполните табл.1, предварительно заменив все прописные символы строчными. Управляющие символы не учитываются. Для определения числа вхождений каждого символа в текст можно воспользоваться опцией «Найти» MS WORDлибо опцией «Заменить», используя для замены символ, не встречающийся в тексте, например, «$». Если символ ни разу не обнаружен в тексте, соответствующая строка таблицы удаляется.

Примечание: в данной работе для упрощения расчетов принимается упрощенная структура текста без различия регистра и без управляющих символов.

2.1.3. По табл.2.1 выполните необходимые расчеты и объясните результаты.

2.2. Составьте эффективные коды Шеннона-Фано и Хаффмана для сжатия выбранного текста.

Определите энтропию и среднее количество двоичных разрядов, необходимых для передачи текста при использовании эффективных кодов.

Проверьте возможность однозначного декодирования полученных кодов, рассмотрев пример передачи слова, состоящего из 6-10 символов.

Примечание: расчеты рекомендуется выполнять в табличной форме, используя MS Excel.

3. СОДЕРЖАНИЕ КОНТРОЛЬНОЙ РАБОТЫ

3.1. Таблицы с расчетами вероятностей вхождения символов в текст и кодирования по алгоритмам Шеннона-Фано и Хаффмана.

3.2. Расчеты энтропии и среднего количества двоичных разрядов, необходимых для передачи текста при использовании эффективных кодов.

3.3. Результаты проверки возможности однозначного декодирования полученных кодов.

3.4. Объяснение результатов и выводы по работе.

3.5. В личный кабинет студент представляет основные результаты работы в виде текстового документа. Для защиты результатов контрольной работы преподавателю представляется файл с расчетами в MS Excel.

Литература:

1. Савельев А.Я. Основы информатики: Учеб. Для вузов.- М.: Изд-во МГТУ им. Н.Э.Баумана, 2001.- 328 с.
2. Темников Ф.Е. и др. Теоретические основы информационной техники.- М.: Энергия, 1979.- 512 с.