RUDN university

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ИНФОРМАТИКИ

Хачумов Вячеслав Михайлович:, д.т.н., проф. г.н.с. ФИЦ ИУ РАН,
зав. лаб. Интеллектуального управления. ИПС РАН.
vmh48@mail.ru , https://icontrol.psiras.ru/

учебный год 2021/2022

RUDN university

Лекция 1. Введение в дисциплину «Теоретические основы информатики"

Понятие информации

Информация (от лат. *informātiō* «разъяснение, представление, понятие о чём-либо»— сведения независимо от формы их представления[1] - сведения, передаваемые людьми устным, письменным или каким-либо другим способом (с помощью условных сигналов, технических средств и т. д.);

В общенаучном понятии, включающее обмен сведениями

- ***Информация***
- - знания о предметах, фактах, идеях и т. д., которыми могут обмениваться люди в рамках конкретного контекста (ISO/IEC 10746-2:1996);
- знания относительно фактов, событий, вещей, идей и понятий, которые в определённом контексте имеют конкретный смысл (ISO/IEC 2382:2015);
- сведения, воспринимаемые человеком и (или) специальными устройствами как отражение фактов материального или духовного мира в процессе коммуникации (ГОСТ 7.0-99).

Клод Шеннон определил *информацию*, как *снятую неопределенность*. Точнее сказать, получение информации - необходимое условие для снятия неопределенности.

Информационная энтропия — мера неопределённости *(Entropy)* некоторой системы, в частности непредсказуемость появления какого-либо символа алфавита. В последнем случае при отсутствии информационных потерь ***энтропия численно равна количеству информации на символ передаваемого сообщения.***

Свойства информации

Значимость. Свойство информации сохранять свою потребительскую ценность для получения в течение времени, то есть не подвергаться моральному старению.

Полезность. Полезность информации оценивается по тем задачам, которые можно решить с ее использованием.

Актуальность. Информация актуальна (своевременна), если она важна в данный момент времени.

Достоверность (правдивость). Информация считается достоверной, если она не противоречит реальной действительности, правильно ее объясняет и подтверждается.

Объективность. Информация может быть объективной или субъективной (зависеть или не зависеть от чьего суждения).

Понятность. Информация понятна, если при ее восприятии нет необходимости в дополнительных сообщениях (не возникает вопросов).

Понятие информатики

Информатика — фундаментальная естественная наука, изучающая общие свойства информации, процессы, методы и средства ее обработки (сбор, хранение, преобразование, перемещение, выдача) **(Ершов А.П.)**
Информатика — наука о способах получения, накопления, хранения, преобразования, передачи, защиты и использования информации

Исторически сложилось так, что исследованием непосредственно информации занимаются две комплексные отрасли науки — кибернетика и информатика.
Информатика, сформировавшаяся как наука в середине XX века, отделилась от кибернетики и занимается исследованиями в области способов получения, хранения, передачи и обработки семантической информации.
Исследования смыслового содержания информации основываются на комплексе научных теорий под общим названием семиотика.

Понятие количества информации

Формулу для вычисления количества информации в случае различных вероятностей событий предложил **К. Шеннон** в 1948 году. В этом случае количество информации определяется по формуле:

$$I = -\sum_{i=1}^{N} p_i \log_2 p_i ,$$

где I - количество информации;

N - количество возможных событий;

p_i - вероятность i-го события.

Эта величина также называется *средней энтропией (Entropy) сообщения*.

Для частного, случая, когда события равновероятны ($p_i = 1/N$), величину количества информации I можно рассчитать по формуле **Р. Хартли**:

$$I = -\sum_{i=1}^{N} \frac{1}{N} \log_2 \frac{1}{N} = \log_2 N .$$

Частота встречаемости букв русского алфавита

Номер по частоте употребления	Буква	Частотность	Частотность %
1	о	0,10983	10,983%
2	е	0,08483	8,483%
3	а	0,07998	7,998%
4	и	0,07367	7,367%
5	н	0,067	6,7%
6	т	0,06318	6,318%
7	с	0,05473	5,473%
8	р	0,04746	4,746%
9	в	0,04533	4,533%
10	л	0,04343	4,343%
11	к	0,03486	3,486%
12	м	0,03203	3,203%
13	д	0,02977	2,977%
14	п	0,02804	2,804%
15	у	0,02615	2,615%
16	я	0,02001	2,001%
17	ы	0,01898	1,898%
18	ь	0,01735	1,735%
19	г	0,01687	1,687%
20	з	0,01641	1,641%
21	б	0,01592	1,592%
22	ч	0,0145	1,45%
23	й	0,01208	1,208%
24	х	0,00966	0,966%
25	ж	0,0094	0,94%
26	ш	0,00718	0,718%
27	ю	0,00639	0,638%
28	ц	0,00486	0,486%
29	щ	0,00361	0,361%
30	э	0,00331	0,331%
31	ф	0,00267	0,267%
32	ъ	0,00037	0,037%
33	ё	0,00013	0,013%

Частота встречаемости букв английского алфавита

Буква	Частота встречаемости
Aa	8,17 %
Bb	1,49 %
Cc	2,78 %
Dd	4,25 %
Ee	12,70 %
Ff	2,23 %
Gg	2,02 %
Hh	6,09 %
Ii	6,97 %
Jj	0,15 %
Kk	0,77 %
Ll	4,03 %
Mm	2,41 %
Nn	6,75 %
Oo	7,51 %
Pp	1,93 %
Qq	0,10 %
Rr	5,99 %
Ss	6,33 %
Tt	9,06 %
Uu	2,76 %
Vv	0,98 %
Ww	2,36 %
Xx	0,15 %
Yy	1,97 %
Zz	0,07 %

Количество информации в сообщении

Простейшей единицей измерения информации является **бит** — единица измерения количества информации, принимающая 2 логических значения: да или нет, истина или ложь, включено или выключено; 1 или 0 в двоичной системе счисления.

Количество информации в сообщении можно подсчитать, умножив количество информации, которое несет один символ, на количество символов.

***Количество информации**, которое содержит сообщение, закодированное с помощью знаковой системы, равно количеству информации, которое несет один знак, умноженному на количество знаков.*

Алгоритм Хаффмана основывается на том, что символы в текстах как правило встречаются с различной частотой. некоторые символы встречаются чаще, а некоторые реже – можно записать часто встречающиеся символы в небольшом количестве бит, а для редко встречающихся символов использовать более длинные коды. Тогда суммарная длина закодированного текста может стать меньше.

Пример: 1)подсчитать количество вхождений каждого символа в тексте «Сжатие Хаффмана». Таблица частот»

С	ж	а	т	и	е		Х	ф	м	н
1	1	4	1	1	1	1	1	2	1	1

Алгоритм Хаффмана

2) Построить дерево

Узел дерева будет образован набором символов и числом - суммарным количеством вхождений данных символов в тексте. Назовем указанное число - весом узла.

Листьями дерева будут узлы, образованные одним символом:

Циклически:

1) Ищем среди узлов, не имеющих родителя, два узла, имеющих в сумме наименьший вес.

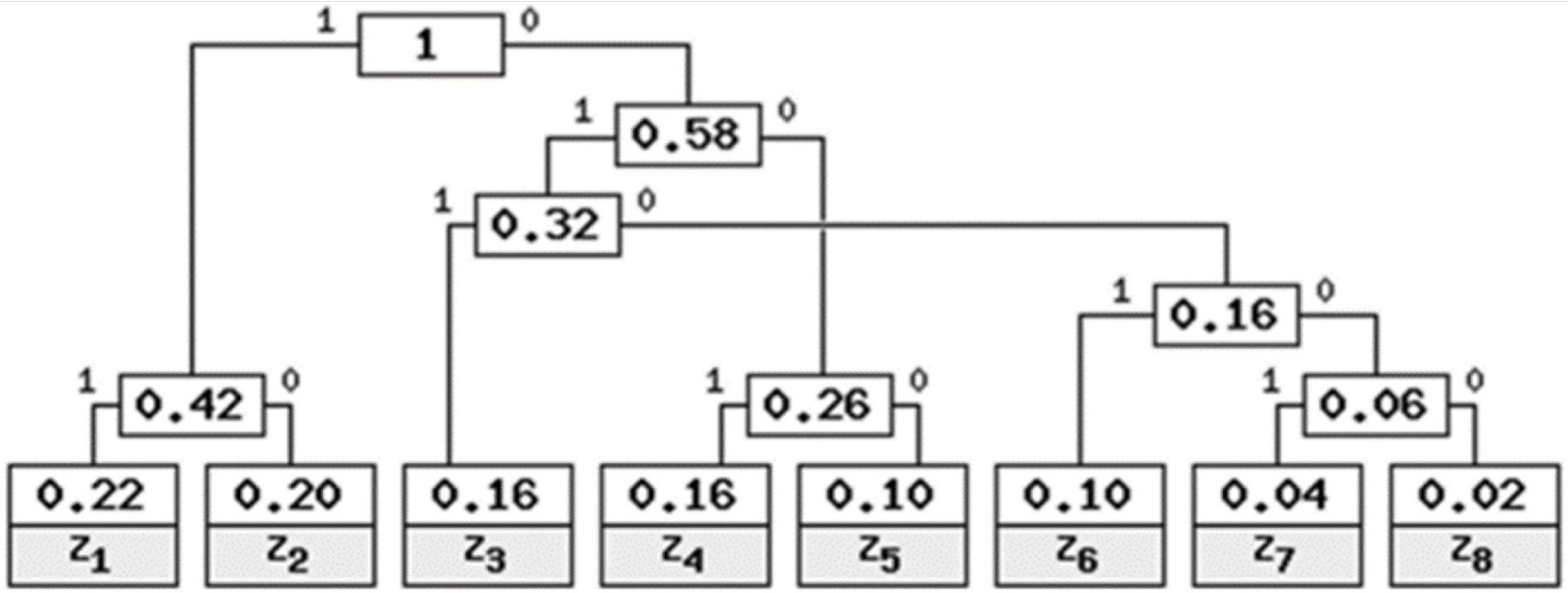
2) Создаем новый узел, его потомками будут два выбранных узла. Его весом будет сумма весов выбранных улов. Его набор символов будет образован в результате объединения наборов символов выбранных узлов.

Для каждого узла, имеющего потомков, пометим дуги, идущие вниз, на одной дуге напишем 0, на другой 1. Чтобы получить код символа, надо, начиная с корня дерева идти вниз до листа соответствующего данному символу.

С	ж	а	т	и	е		Х	а	ф	ф	м	а	н	а
0001	0000	01	0010	0011	1011	1010	111	01	110	110	1000	01	1001	01

Алгоритм Хаффмана. Пример построения дерева

Буква	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8
Вероятность	0.22	0.20	0.16	0.16	0.10	0.10	0.04	0.02



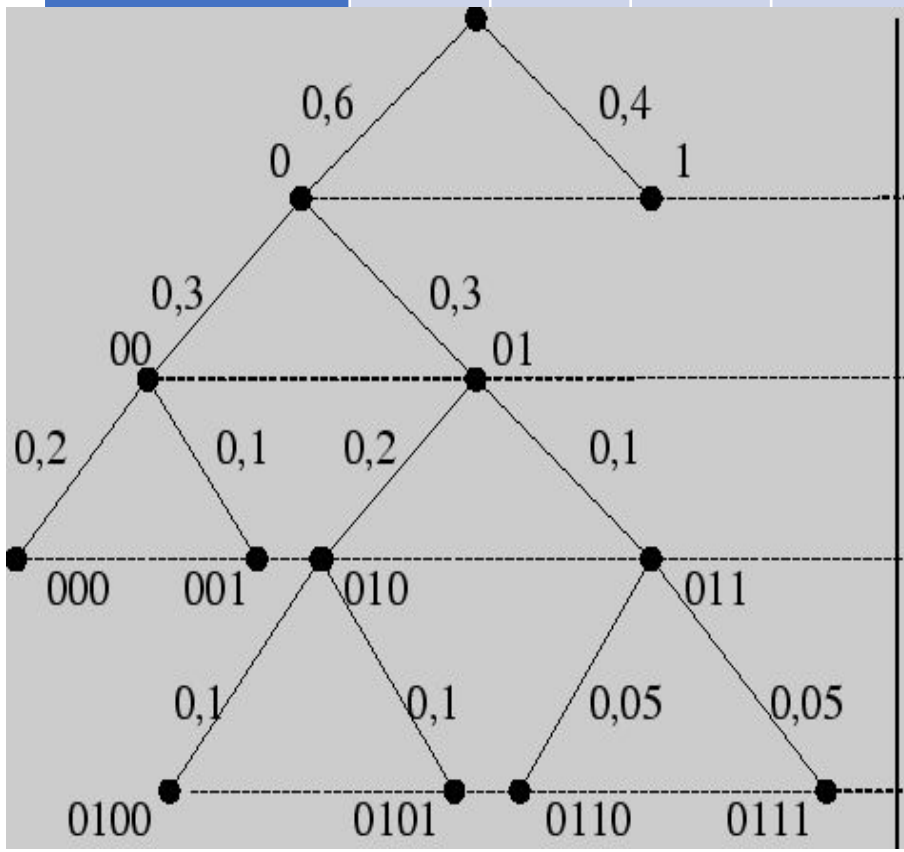
Буква	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8
Код	11	10	011	001	000	0101	01001	01000

Код Фано

Все сообщения записываются в таблицу по степени убывания вероятности и разбиваются на две группы примерно (насколько это возможно) равной вероятности.
Соответственно этой процедуре из корня кодового дерева исходят два ребра, которым в качестве весов присваиваются полученные вероятности. Двум образовавшимся вершинам приписывают кодовые символы 0 и 1. Затем каждая из групп вероятностей вновь делится на две подгруппы примерно равной вероятности. В результате многократного повторения процедуры разделения вероятностей и образования вершин приходим к ситуации, когда в качестве веса, приписанного ребру бинарного дерева, выступает вероятность одного из данных сообщений. В этом случае вновь образованная вершина оказывается листом дерева, т.к. процесс деления вероятностей для нее завершен. Задача кодирования считается решенной, когда на всех ветвях кодового бинарного дерева образуются листья.

Код Фано

сообщение	1	2	3	4	5	6	7
вероятность	0,4	0,2	0,1	0,1	0,1	0,05	0,05



Коды	Сообщения
1	1
000	2
001	3
0100	4
0101	5
0110	6
0111	7

Цена кодирования (средняя длина кодового слова является критерием степени оптимальности кодирования. Это сумма произведений длины кода на вероятность появления кода. $L = \sum_{i=1}^{7} l_i p_i$ =1·0.4+3·0.2+3·0.1+4·0.1+4·0.1+4·0.05+4·0.05=2.5

Оценка информационной значимости признаков (ак. Журавлев Ю.И.)

Пусть задана таблица бинарных признаков для случая двух классов.

Определение 1: тестом таблицы называется совокупность столбцов таких, что после удаления из таблицы всех прочих столбцов все пары строк, принадлежащие разным классам, различны.

Определение 2: тест называется тупиковым, если никакая его истинная часть не является тестом.

Таким образом, тест – это минимальный набор признаков, для которого отсутствуют одинаковые векторы-строки в разных классах обучающей выборки.

Тупиковый тест - это не сжимаемое далее описание, которое содержит все сведения о разделении исходного множества объектов на классы.

Тупиковый тест не содержит внутри себя другие тесты с меньшим числом признаков.

Оценка информативности признаков

Если признак k участвует в формировании n_k тупиковых тестов, то его значимость определяется по формуле $\gamma_k = \frac{n_k}{N}$, где N – общее количество тупиковых тестов.

Таблица. Объекты и их бинарные признаки

Объекты (ситуации)	x_1	x_2	x_3	x_4	x_5	x_6	Класс
a_1	1	1	0	1	1	0	Ω_1
a_2	0	0	0	0	0	1	
a_3	1	0	1	0	1	0	
a_4	0	1	1	0	1	1	Ω_2
a_5	1	0	0	0	1	0	

1. Выписываем все тупиковые тесты:

$$x_1x_2x_3,\ \ x_1x_3x_4,\ \ x_2x_3x_5,\ \ x_2x_3x_6,\ \ x_3x_4x_6,$$

2. Находим значимость каждого признака:

$$\gamma_1 = \frac{2}{5},\ \ \gamma_2 = \frac{3}{5},\ \ \gamma_3 = \frac{5}{5} = 1,\ \ \gamma_4 = \frac{2}{5},\ \ \gamma_5 = \frac{1}{5},\ \ \gamma_6 = \frac{2}{5}.$$

Чем больше γ_k, тем важнее признак k для описания допустимых объектов.

Задания на самостоятельную работу

- 1.Выполнить кодирование по Хаффману и по Фано произвольного короткого текстового сообщения (с повторяющимися буквами)
- 2.Определить количество информации в русском алфавите по Шеннону
- 3. Определить тесты для заданной таблицы с учебной выборкой по методу академика Ю.И.Журавлева.

Номер ситуации	Признаки												Класс
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	
1	0	1	0	0	0	1	1	1	0	0	1	1	1
2	0	1	1	1	0	0	1	0	1	0	1	1	
3	0	1	0	1	1	1	0	1	0	0	0	0	
4	0	0	1	0	1	0	1	0	1	1	1	1	
5	1	1	0	1	0	1	1	0	1	0	1	0	
6	1	0	1	0	1	1	0	1	0	1	0	1	0
7	1	0	1	0	1	1	1	0	0	1	0	0	
8	1	0	0	1	1	1	0	1	0	0	0	0	
9	1	0	1	1	1	1	0	0	1	0	0	1	
10	0	1	1	0	0	0	0	1	0	1	1	1	

БЛАГОДАРЮ ЗА ВНИМАНИЕ!

vmh48@mail.ru