Информация и ценность

Краткое содержание

Количественная теория измеряет информацию через число состояний и их вероятности, а качественная — через достоверность, полезность и влияние информации на достижение цели.

Формула Хартли

Информация в символе

Пусть алфавит содержит N символов, а сообщение имеет длину k. Количество информации в одном символе обозначается i и определяется формулой:

$$\ni = \log_2 N$$

Количество информации во всём сообщении:

$$I = k \log_2 N = ki$$

Отсюда мощность алфавита:

$$N = 2^i$$

Если алфавит содержит N символов, то число различных сообщений длины k равно:

$$N^k$$

Пример с ДНК

ДНК строится из **4 нуклеотидов**, поэтому мощность её алфавита равна $N = 4$. Информация в одном нуклеотиде:

$$\ni = \log_2 4 = 2 \text{ бита}$$

Для простейшего вируса с цепочкой примерно из 5000 нуклеотидов число возможных вариантов составляет:

$$4^{5000}$$

Для человеческой ДНК, содержащей около 10^9 нуклеотидов, разнообразие возможных последовательностей оценивается как:

$$4^{10^9}$$

Даже небольшие изменения последовательности ДНК создают огромное число вариантов мутаций.

Ограничение модели

Формула Хартли предполагает, что все символы или состояния:

- равновероятны;
- содержат одинаковое количество информации;
- появляются независимо от различий в вероятностях.

Это упрощение стало главным недостатком формулы Хартли.

Энтропия Шеннона

В **1948 году** Клод Шеннон предложил вероятностный подход к измерению информации. Состояния объекта могут иметь разные вероятности, поэтому информация зависит не только от числа вариантов, но и от вероятности каждого варианта.

> *Энтропия Шеннона — среднее количество информации, приходящееся на одно событие или состояние.*

Формула Шеннона:

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

Здесь:

- N — количество возможных состояний;
- i — номер слагаемого;
- p_i — вероятность i-го состояния;
- I — среднее количество информации.

Вероятности должны образовывать **полную группу событий**:

$$\sum_{i=1}^{N} p_i = 1$$

Каждая вероятность находится в диапазоне $0 \le p_i \le 1$.

Знак минус необходим потому, что для $0 < p_i < 1$ значение $\log_2 p_i$ отрицательно. Минус перед суммой делает результат положительным.

При $p_i = 0$ выражение $p_i \log_2 p_i$ принимают равным нулю: первый множитель обнуляет слагаемое.

Связь с формулой Хартли

Если все N состояний равновероятны, то:

$$p_i = \frac{1}{N}$$

Тогда формула Шеннона преобразуется в формулу Хартли:

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

Следовательно, **формула Хартли является частным случаем формулы Шеннона**.

Единицы измерения

Единица зависит от основания логарифма:

Основание логарифма	**Единица информации**
2	бит
10	дит, или хартли
e	нат

Бит стал основной единицей благодаря распространению компьютеров и двоичной системы . Натуральный логарифм широко используется в инженерных дисциплинах, связанных с непрерывными процессами и экспонентой.

Свойства энтропии

Чем больше число возможных состояний N, тем больше потенциальное количество информации. Если состояние заранее известно, информация равна нулю: одна вероятность равна единице, остальные — нулю.

$$p_1 = 1, \quad p_2 = \cdots = p_N = 0 \quad \Rightarrow \quad I = 0$$

Максимальная информация при фиксированном N получается тогда, когда все состояния **равновероятны**: неопределённость максимальна, поэтому получение результата сообщает больше всего нового.

Пример энтропии

В ящике находятся 20 шаров:

- 10 чёрных;
- 4 белых;
- 6 красных.

Возможны три состояния, но они не равновероятны:

$$p_{ч} = \frac{10}{20} = \frac{1}{2}, \qquad p_{б} = \frac{4}{20} = \frac{1}{5}, \qquad p_{к} = \frac{6}{20} = \frac{3}{10}$$

Количество информации о цвете вынутого шара:

$$I = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{5}\log_2\frac{1}{5} + \frac{3}{10}\log_2\frac{3}{10}\right)$$

Приближённо:

$$I \approx 1{,}49 \text{ бита}$$

Использовать формулу Хартли здесь нельзя напрямую, поскольку цвета имеют разные вероятности. Формула $\log_2 3$ дала бы значение только для трёх равновероятных цветов.

Качество информации

Количество информации не показывает, насколько сведения полезны. Для качественной оценки важны три параметра.

Достоверность

Факт считается более достоверным, если его подтверждают **не менее трёх независимых источников**. Источники не должны просто пересказывать один и тот же материал.

Доверие повышают:

- квалификация и осведомлённость автора;
- независимое подтверждение;
- умение отличать факт от мнения;
- повторная проверка непонятных или важных утверждений.

Для научной работы нельзя ограничиваться одним-двумя источниками. Полезно сопоставлять несколько учебников, статей и монографий: разные авторы могут по-разному трактовать один вопрос.

Полнота

Полная информация создаёт основу для правильного решения. Поиск обычно проходит по уровням:

1. **Источники общих сведений** — энциклопедии, справочники, научно-популярные статьи.
2. **Специальные источники** — специализированные энциклопедии и статьи.
3. **Монографии и научные труды** — глубокое изучение проблемы.
4. **Дополнительные материалы** — новые исследования, архивы, экономические и политические данные, современные примеры.

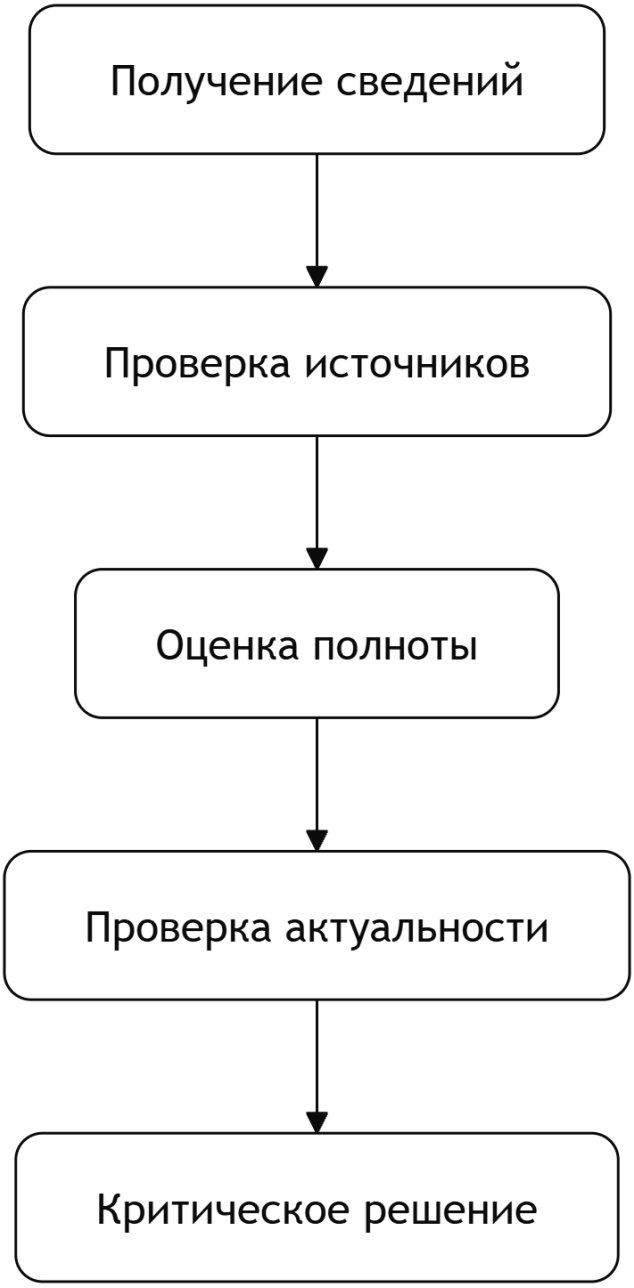
Остановка на поверхностных сведениях снижает качество работы.

Актуальность

Актуальная информация соответствует современному состоянию вопроса и времени, в котором принимается решение. В IT знания быстро устаревают, поэтому специалисту необходимо постоянно обновлять компетенции.

Сделано на сайте Зачёт | zachet.app

Для современных прикладных тем обычно проверяют год публикации; статьи старше примерно пяти лет могут быть неактуальны. Для фундаментальных вопросов допустимы более старые источники.



Ценность по Харкевичу

Александр Александрович Харкевич предложил оценивать не количество, а **ценность информации** через изменение вероятности достижения цели.

$$I_ц = \log_2 \frac{P_1}{P_0}$$

Здесь:

- P_0 — вероятность достижения цели до получения информации;

- P_1 — вероятность достижения цели после получения и использования информации.

Ценность информации зависит от цели конкретного субъекта, поэтому она **субъективна**: одни и те же сведения могут быть крайне важны одному человеку и бесполезны другому.

Три результата получения информации

Соотношение вероятностей	Результат
$P_1 > P_0$	полезная информация, приближающая к цели
$P_1 = P_0$	информационный шум, ничего не изменивший
$P_1 < P_0$	дезинформация, отдаляющая от цели

Дезинформация — введение в заблуждение с помощью искажённых, неполных или устаревших сведений. Она может использоваться для манипулирования, чтобы другая сторона действовала в интересах манипулятора. Примеры сфер применения — война, политика, реклама и конкуренция.

Информационный шум — данные, которые не соответствуют информационной потребности субъекта, не помогают достичь цели или не содержат новизны.

Информация как ресурс

Для достижения цели используются три типа ресурсов:

- вещество;
- энергия;
- информация.

Информация помогает экономить вещественные и энергетические ресурсы: выбрать более короткий путь, правильно организовать работу, распределить людей по задачам, снизить затраты и избежать препятствий.

Если R_0 — ресурсы без использования информации, а R_1 — ресурсы после её применения, то экономия ресурсов:

$$\Delta R = R_0 - R_1$$

- $\Delta R > 0$ — информация полезна и уменьшает затраты;
- $\Delta R = 0$ — информация не повлияла на результат;
- $\Delta R < 0$ — информация привела к дополнительным затратам, например из-за дезинформации.

Вещественные и энергетические ресурсы истощаются, а информационный ресурс обычно сохраняется после использования и может применяться другими людьми.

Если после получения информации $P_1 = 1$, цель достигнута достоверно. Тогда:

$$I_{ц} = \log_2 \frac{1}{P_0} = -\log_2 P_0$$

Ценность равна количеству неопределённости, снятой получением информации.

Сделано на сайте Зачёт | zachet.app

Формула Харкевича стала важным шагом к оценке качества информации, но отождествлять **количество** и **ценность** нельзя: это разные характеристики, даже если они выражаются через логарифм и измеряются в похожих единицах.

Сделано на сайте Зачёт | zachet.app