

Teoria informacji

Autorzy:

Maciej Jurczyk

Daniel Kosytorz

Informacja

- ▶ Wyobraźcie sobie, że chcecie podzielić się z kimś swoją myślą. Jest wiele sposobów aby to zrobić. Możemy to narysować, napisać, wyryć na czymś, wysłać maila czy po prostu powiedzieć. Co łączy te wszystkie sposoby? Język. Może on przyjmować różne formy np. sygnały dźwiękowe, symbole czy czyny. Wszystko to prowadzi do jednego: informacji.
- ▶ Informacja jest tym, co pozwala jednemu umysłowi wpływać na drugi. Informację, niezależnie od jej postaci, mierzymy w podstawowych jednostkach, tak, jak mierzymy masę różnych rzeczy przy użyciu standardowej miary. Ich miarę nazywamy entropią, o której więcej będzie później. Wyobraźcie sobie wagę do ważenia informacji. Intuicyjnie wiemy, że jedna strona z nieznanej nam książki zawiera mniej informacji niż cała książka. Możemy dokładnie określić, ile, stosując jednostkę zwaną bitem. Bit wyraża coś bardzo prostego: odpowiedź na pytanie typu: „tak czy nie”.

Informacja w starożytności

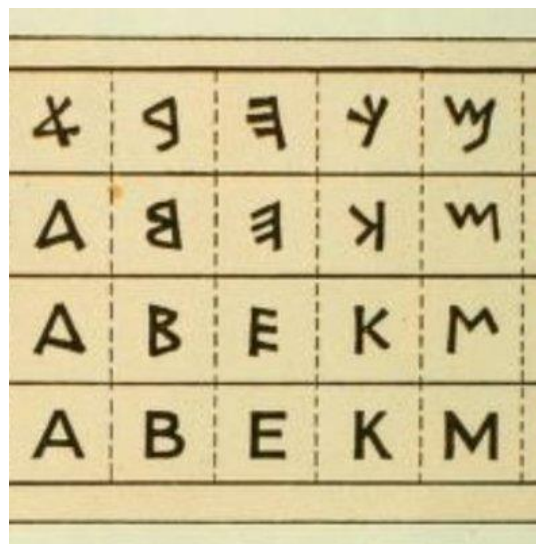
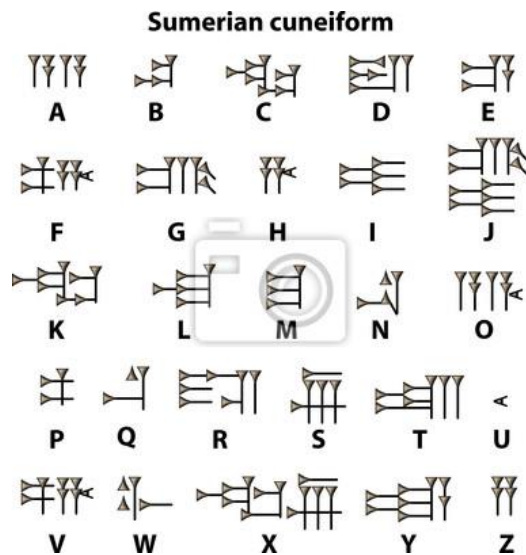
Piktogramy i hieroglify

- ▶ Piktogram, czyli rysunek podobny do przedmiotu, który przedstawia. Jedne z najstarszych dowodów łączenia symboli znaleziono w starożytnej Mezopotamii, gdzie żyli Sumerowie. Zachowały się tu gliniane tabliczki, jedne z najstarszych znalezionych pisanych dokumentów.
- ▶ Na prostokątnych tabliczkach zapisano ceny zwierząt, ich liczbę na pastwisku oraz składane z nich ofiary. Zamiast rysować 10 owiec, użyli symbolu oznaczającego liczbę 10, z małych wgłębień, do którego dodali symbol owcy lub osła. Razem oznacza to po prostu 10 owiec. Coś takiego nazywamy protopismem.



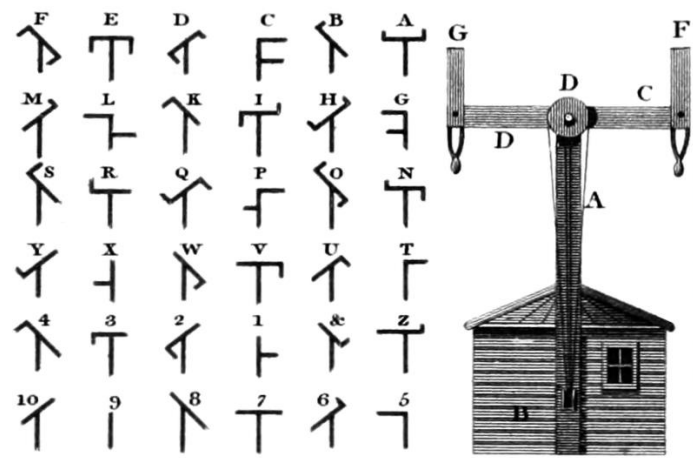
Alfabet

- ▶ W starożytnym Egipcie najpierw używano hieroglifów, które zamieniono z czasem w pismo hieratyczne, ponieważ ludzie częściej korzystali z papiirusu, symbole stopniowo ewoluowały, by dało się zapisywać je szybciej. Symbole opierały się na hieroglifach, ale były uproszczone, dla większej szybkości pisania.
- ▶ W Mezopotamii ok 3000 r.p.n.e stosowano pismo klinowe w celach podatkowych.
- ▶ Ok. 1000 r. p. n. e. zaczął szerzyć się alfabet fenicki. Pojawił się nad Morzem Śródziemnym. Zawierał on tylko 22 symbole. Symbole reprezentujące te dźwięki często były zapożyczone z hieroglifów; nazwa litery wywodziła się od kojarzonego z hieroglifem dźwięku. Np. „mem” czyli „woda”, stało się znaną nam obecnie literą „m”. Słowo „alf”, czyli „wół”, przerodziło się w literę „a”. Ale największa moc tego alfabetu, której nie uświadamiali sobie jego wynalazcy, polegała na tym, że nie ograniczał się do języka semickiego. Wystarczyły nieznaczne zmiany, by te cudowne litery dopasowały się do różnych języków Europy, Indii i Azji. Alfabet ten przyczynił się do szerzenia pisma na świecie. Był źródłem alfabetu greckiego, a później rzymskiego, który znamy dziś.



Telegraf optyczny

- ▶ Jednym z najstarszych pomysłów przekazywania wiadomości na duże odległości, a znanym już w starożytności, był telegraf optyczny. Wówczas używali go Grecy i Kartagińczycy. Telegraf taki składał się z pochodni oraz zegara wodnego. Pochodnią sygnalizowano uruchomienie i wyłączenie zegara, a poziom wody w nim oznaczał określoną wiadomość. Inną formą telegrafu optycznego stosowanego przez Greków były odpowiednio ustawiane pochodnie. W ten sposób przekazywali wiadomości przy użyciu liter, bo każda z nich miała przyporządkowany sobie odpowiedni układ świateł.
- ▶ W 1791 roku Francuz Claude Chappe stworzył semaforowy telegraf optyczny. Zasada jego działania oparta była na odpowiednim ustawieniu belek sygnałowych, umieszczonych na maszcie lub jakiejś wyniosłości terenu. Do przekazywania wiadomości na duże odległości stosowano system stacji przekaźnikowych. Pierwszą linią tego typu telegrafu było połączenie Paryża z Lille. Również w Polsce zastosowano ten wynalazek łącząc Warszawę z Modlinem, Petersburgiem i Moskwą.



Telegraf elektryczny

- ▶ Od 1832 roku pracował nad telegrafem elektrycznym również amerykański wynalazca Samuel Morse. W 1835 roku zbudował on prototyp swego telegrafu, a 2 września 1837 roku zademonstrował go w Nowym Jorku i otrzymał nań patent (28 IX 1837 rok). W 1840 roku Morse wprowadził stosowany do dzisiaj kod telegraficzny złożony z kropek i kresek (krótkie i długie impulsy) - tzw. alfabet Morse'a.



Alfabet Morse'a

- ▶ Samuel Morse opatentował rozwiązanie oparte na pomysł, że prąd płynie lub nie, a przerwy można ustawić tak, by tworzyły sens. Jego rozwiązanie doprowadziło do powstania legendarnego elementu: prostej dźwigni ze sprężyną, czyli klucza, do wciskania palcem. Na drugim końcu znajdowała się dźwignia ze sprężyną, podciągana i opuszczana przez silny elektromagnes. Wciśnięcie klucza na krótko nazwano kropką. Wciśnięcie klucza na trzy jednostki czasu nazywało się kreską. Wszystko - w równych odległościach.
- ▶ Krótsze sekwencje symboli przypisano częściej występującym literom. Co zdumiewające, właśnie prostota systemu z kluczem sprawiła, że był szybszy od wszelkich przycisków i dźwigni w europejskich telegrafach igłowych. Szybkość przekazu wzrosła do ponad 135 liter na minutę. 24 maja 1844 r. nadano pierwszy przekaz. Brzmiał: „Co Bóg uczyni?”. Nazajutrz w „New York Tribune” napisano: „Zwolna dokonuje się cud unicestwienia przestrzeni”. Pamiętajmy, że wówczas 90% wiadomości nadal rozwożono konno. Wszystko, co dotyczyło informacji, opierało się teraz na telegrafie i kodzie Morse'a.
- ▶ Do 1900 r. ceny spadły do 30 centów za wiadomość, a w tamtym roku nadano ponad 63,2 mln wiadomości. Zaczynając używać tego systemu, ludzie zastanawiali się, jak trochę oszczędzić. Powstały popularne książki kodów, wiążące słowa z częstymi zdaniami. Np. „Blade” oznaczało: „Proszę zarezerwować dla mnie i dla mojej rodziny to, co następuje”.

Co to za słowo?

A	• -	M	- -	Y	- • - -
B	- • • •	N	- •	Z	- - • •
C	- • - •	O	- - -	1	• - - - -
D	- • •	P	• - - •	2	• • - -
E	•	Q	- - • -	3	• • • - -
F	• • - •	R	• - •	4	• • • -
G	- - •	S	• • •	5	• • • • •
H	• • • •	T	-	6	- • • • •
I	• •	U	• • -	7	- - • • •
J	• - - -	V	• • • -	8	- - - • •
K	- • -	W	• - -	9	- - - - •
L	• - • •	X	- • • -	0	- - - - -

• • - • • • - • - - • - • - - • - • - - • -

Współczesna teoria informacji

- ▶ Współcześnie dążymy do przyspieszania przesyłu danych przez: przepustowość i kompresję danych
- ▶ Rozpoczyna się w I połowie XX wieku
- ▶ Podobnie jak wcześniej chciano przyspieszyć przekaz informacji tak aby był jasny sam w sobie. Wykorzystywano jednak coraz to nowsze osiągnięcia ludzkiego umysłu co zdecydowanie ułatwiało te dążenia. Współcześnie zatem przyspieszono przesyłu danych. Szczególnie po przez:
 - Rozważania na temat przepustowości sieci
 - Kompresję danych przesyłanych

Przepustowość

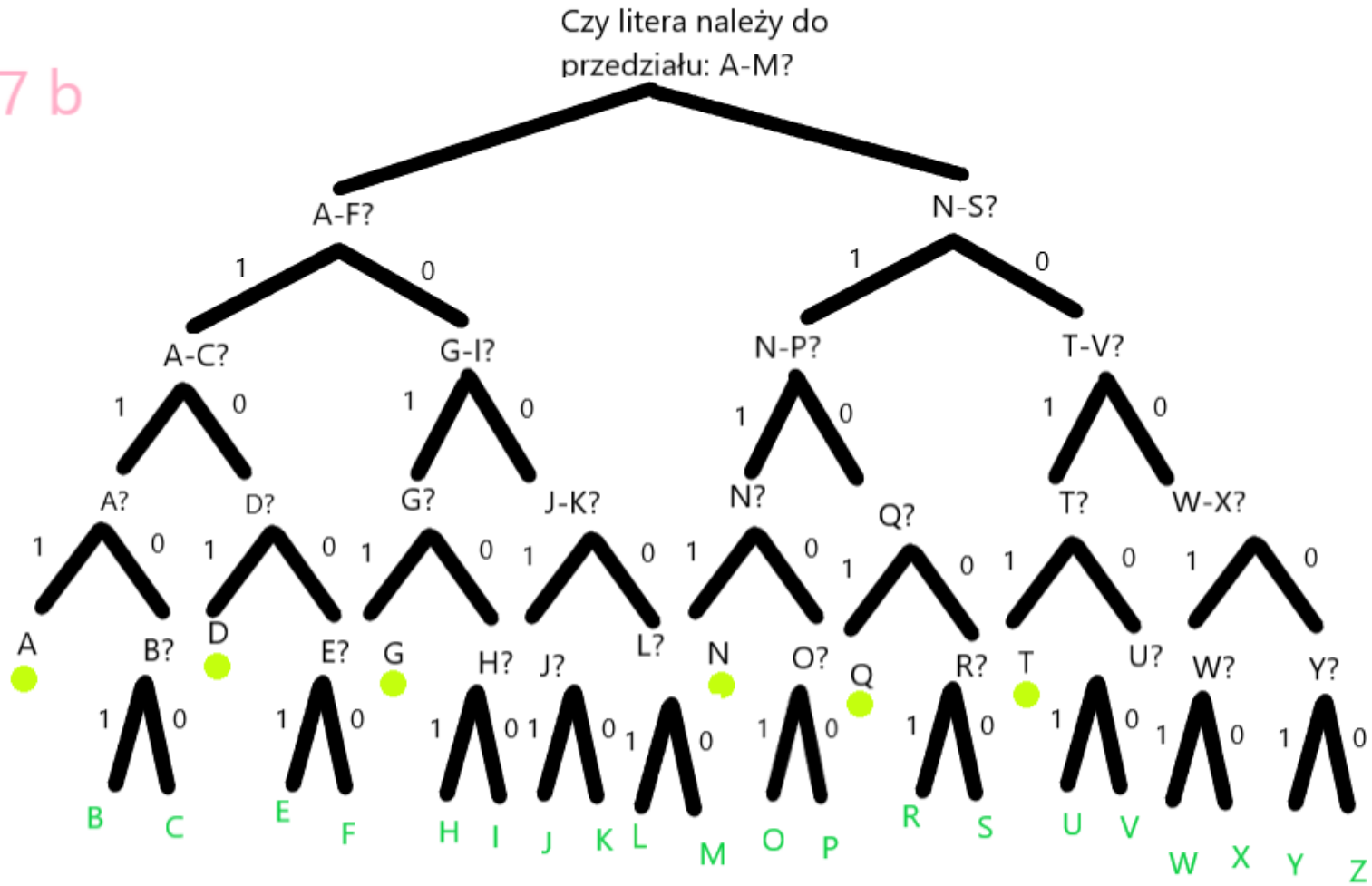
- ▶ Aby zwiększyć przepustowość wbudowanych linii przesyłowych wykorzystywano cztery rodzaje napięcia: -3 V, -1 V, 1 V oraz 3 V.
- ▶ Starano się również wysłać wiadomości jak najszybciej to możliwe, jednakże przy odpowiednio szybkim wprowadzaniu impulsów, na większym dystansie wysyłane fale zlewały się co często uniemożliwiało prawidłowe odczytanie wiadomości.
- ▶ Elementem mocno związanym z przepustowością jest liczba rodzaj symboli lub znaków użytych do przysłania wiadomości. Im większe ich urozmaicenie tym więcej jesteśmy w stanie przekazać informacji w tym samym czasie.
- ▶ Odpowiada temu wzór S^N : gdzie S to liczba symboli, a N to ilość przesyłanych symboli na sekundę.
- ▶ Np. w przekazie opisanym wyżej (różne napięcia elektryczne), przyjmując możliwość przesyłania 2 impulsów na sekundę możemy przekazać prawie 1000 znaków w ciągu minuty.

Przenosząc przepustowość na konkretne zagadnienie związane z informatyką...

Ile bitów potrzeba, aby
określić dowolną literę
wskazaną przez osobę?

Avg 4,7 b

A, B, C, D,
E, F, G, H,
I, J, K, L,
M,
N,
O, P, Q, R,
S, T, U, V,
W, X, Y, Z

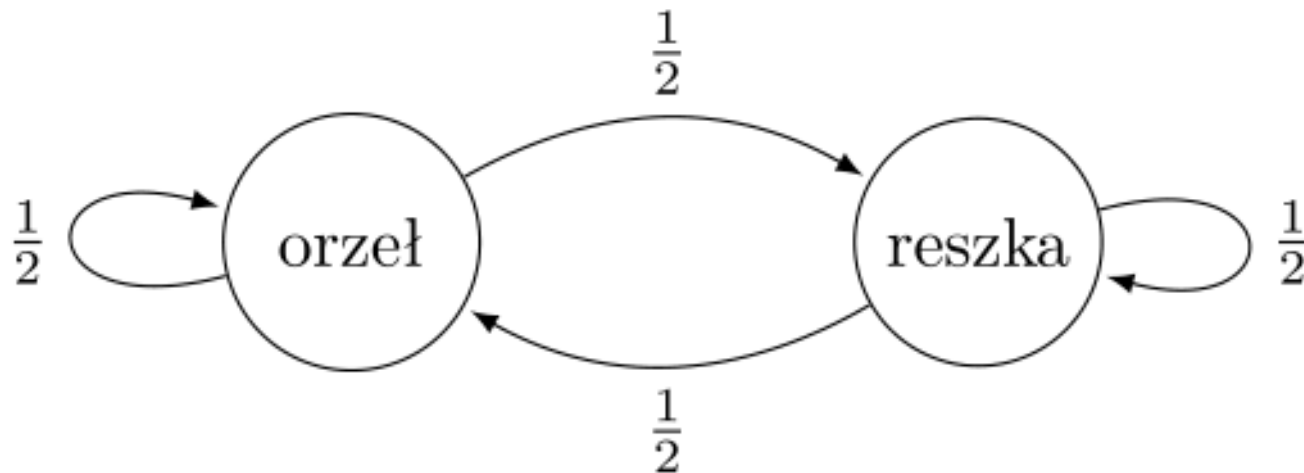


Prawdopodobieństwo a kompresja danych

Bardzo pomocne w kompresji danych okazało się prawdopodobieństwo.

Andriej Markow w 1906 r. opracował podstawy, które później udoskonalone pozwoliły na utworzenie łańcuchu Markowa. Jest to ciąg zdarzeń losowych, w których wynik poprzedniego losowania warunkuje wynik losowania następującego po nim.

Dzisiaj schemat ten wykorzystuje m. in. firma Google w sortowaniu wyszukiwanych stron, czyli w PageRank'u



Kompresja danych

Cały czas dążono do kompresji danych. Jej częścią jest skracanie kodu równoznacznego z danym znakiem do minimum. Z pomocą prawdopodobieństwa i łańcuchu Markowa udaje się to w znacznym stopniu. Wykorzystując zależność wysokości prawdopodobieństwa stwierdzono, że można do zapisu najczęściej występujących informacji wykorzystać mniej bitów, a do rzadziej występujących więcej bitów.

I tak na przykład:

Przyjmijmy, że w alfabecie od A do D prawdopodobieństwo kolejnych liter wynosi:
A- 50% , D - 25 % oraz B i C po 12,5%.

Wykorzystując tą informację możemy skompresować dane potrzebne do ich przekazu w np.. w systemie binarnym dokładnie tak:

A- 1 B- 011 C- 010 D- 00 *

* Musimy jednak uważać na powtarzające się znaki, gdyż kod 011 przy oznaczeniu litery D = 0, mógłby oznaczać jednocześnie= „B” = „DAA”, dlatego ważne jest tworzenie unikalnych kodów

Entropia

Zależność tą opisał m.in. Claude Shannon. W 1948 r. w dziele „Matematyczna teoria informacji” opisał podwaliny pod pojęcie dzisiaj rozumiane jako teorię informacji. Opisując kompresję danych stworzył wzór na średnią liczbę bitów potrzebnych do zapisu danych. Nazywa się to Entropią. Shannon przebadął tą zależność również na języku angielskim rozpatrując poszczególne ciągi liter (jedno oraz kilku literowe), a także na ciągu pojedynczych oraz kilku słów.

Wzór na obliczanie entropii:

$$H(X) = \sum_{i=1}^n p(x_i) \log_r \frac{1}{p(x_i)} = - \sum_{i=1}^n p(x_i) \log_r p(x_i)$$

$H(X)$ - entropia

$p(x)$ - prawdopodobieństwo wystąpienia konkretnej zmiennej

r - liczba zmiennych (np. w systemie binarnym dwie)

Dziękuję za uwagę 😊

- ▶ Autorzy:
- ▶ Maciej Jurczyk
- ▶ Daniel Kosytorz