# Kody Huffmana oraz entropia przestrzeni produktowej
# Kod Shannona-Fano oraz Entopia względna i warunkowa

Zuzanna Kalicińska
Piotr Góra

27 maja 2004

## 1 Optymalny kod bezprefiksowy

**Definicja 1.** *Kod nad alfabetem* $\{0,1\}$*, w którym reprezentacja żadnego znaku nie jest prefiksem reprezentacji innego znaku, nazywamy binarnym kodem bezprefiksowym.*

Za pomocą kodu bezprefiksowego można uzyskać maksymalny stopień kompresji osiągalny za pomocą kodów przypisanych znakom na stałe.

Binarny kod bezprefiksowy można reprezentować jako drzewo binarne, gdzie liście odpowiadają elementom $s \in S$, a gałęzie symbolom 0 i 1 (zależnie od potomka, do którego prowadzą). Ścieżka od korzenia do liścia reprezentuje słowo kodujące element znajdujący się w tym liściu.

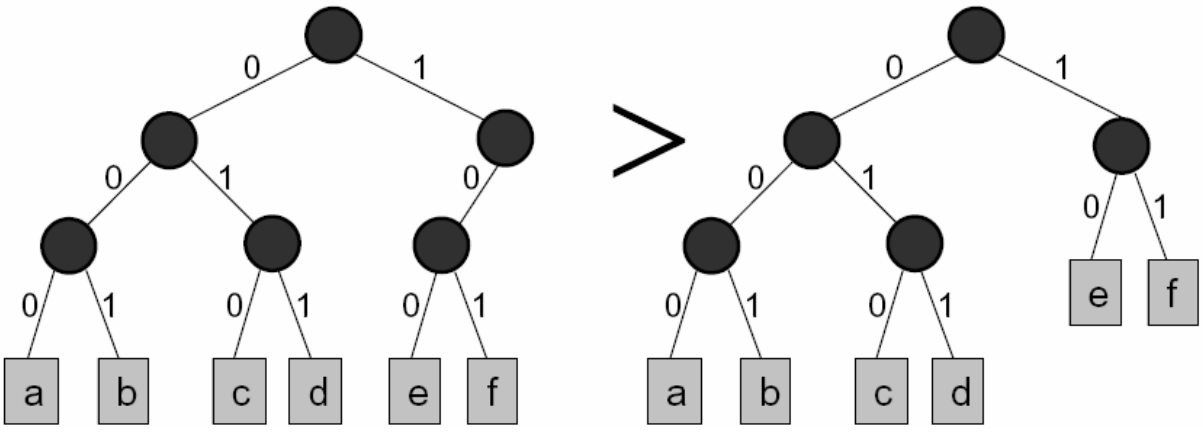
Przypomnijmy z pierwszego wykładu, iż optymalność kodu jest szacowana przez jego długość

$$L(\varphi) = \sum_{s \in S} p(s) \cdot |\varphi(s)|$$

interesuje nas oczywiście najkrótszy kod dla danych $\langle S, p \rangle$

$$L_k(S) = \min\{L(\varphi) : \varphi \ \ to\ k-arny\ bezprefiksowy\ kod\ na\ \langle S, p \rangle\}$$

**Uwaga 1.** *Drzewo odpowiadające optymalnemu kodowi musi być pełne.*
Takie drzewo ma dokładnie $|S|$ liści i $|S|-1$ węzłów wewnętrznych.




Rysunek 1: Struktura drzewa odzwierciedla między innymi optymalność kodu.

**Dowód.** Jeśli w drzewie istnieje węzeł o stopniu 1, to można zastąpić ten węzeł jego potomkiem i w ten sposób uzyskać kod mniejszej długości.

## 2 Kody Huffmana

Algorytm Huffmana polega na przypisywaniu skończonemu zbiorowi symboli kodów o zmiennej liczbie bitów. Idea jest taka, że symbolom o większym prawdopodobieństwie występowania przypisujemy krótsze kody.

**Algorytm Huffmana** dla zbioru $\langle S, p \rangle$

- Jeśli $|S|=1$, $S=\{s\}$ to $\varphi_S^H(S)=\varepsilon$ (dopuszczamy, że kod może przyjąć $\varepsilon$, ale tylko gdy $|S|=1$ - założenie jedynie do dowodu optymalności)
- Jeśli $|S|\geq 2$, to niech $q_0, q_1$ będzie parą o najmniejszej wartości $p(q_0)+p(q_1)$.
  Następnie łączymy $q_0, q_1$ w jeden element i oznaczamy jako $q_0 \vee q_1$
  $$S'=(S-\{q_0,q_1\})\cup\{q_0\vee q_1\},$$
  którego prawdopodobieństwo wynosi
  $$p(q_0\vee q_1)=p(q_0)+p(q_1)$$
  Wreszcie określamy kod
  $$\varphi_S^H(s)=\varphi_{S'}^H(s) \text{ dla } s\neq q_0,q_1$$
  $$\varphi_S^H(q_i)=\varphi_{S'}^H(q_0\vee q_1)i \quad \text{dla i = 0, 1 (przedłużenie kodu o bit } i\text{)}$$

**Przykład 1.**

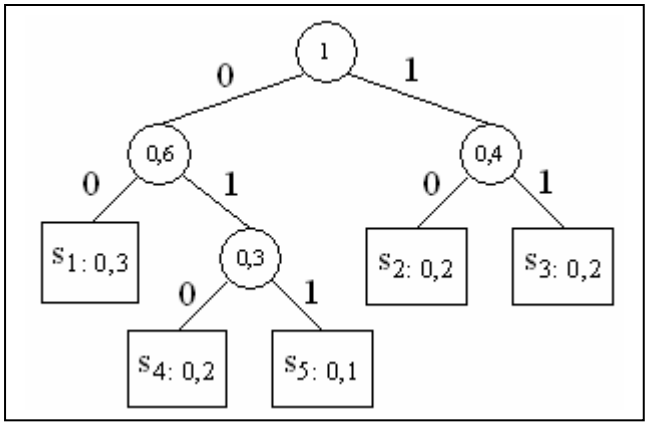
$$S=\{s_1,s_2,s_3,s_4,s_5\},\quad p(s_1)=0,3\quad p(s_2)=0,2\quad p(s_3)=0,2\quad p(s_4)=0,2\quad p(s_5)=0,1$$

$$S'=\{s_1,s_2,s_3,s_4\vee s_5\},\quad p(s_1)=0,3\quad p(s_2)=0,2\quad p(s_3)=0,2\quad p(s_4\vee s_5)=0,3$$

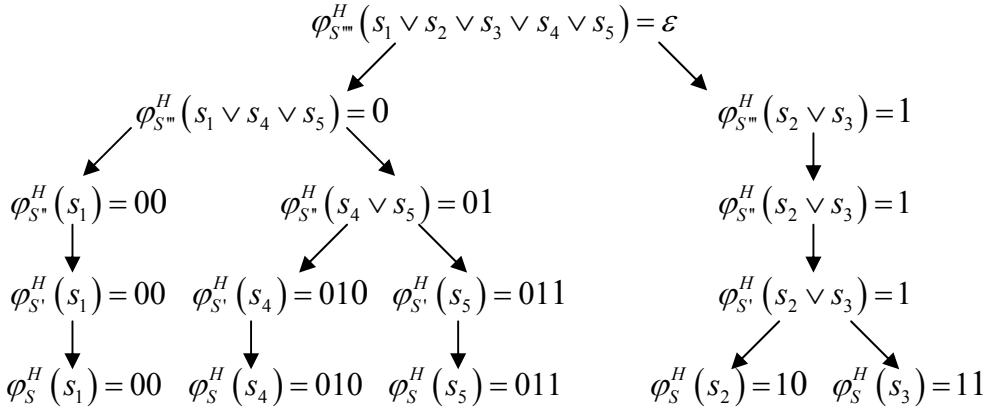
$$S''=\{s_1,s_2\vee s_3,s_4\vee s_5\},\quad p(s_1)=0,3\quad p(s_2\vee s_3)=0,4\quad p(s_4\vee s_5)=0,3$$

$$S'''=\{s_1\vee s_4\vee s_5,s_2\vee s_3\},\quad p(s_1\vee s_4\vee s_5)=0,6\quad p(s_2\vee s_3)=0,4$$

$$S''''=\{s_1\vee s_2\vee s_3\vee s_4\vee s_5\},\quad p(s_1\vee s_2\vee s_3\vee s_4\vee s_5)=1$$



Rysunek 2: Drzewo odpowiadające kodowi z przykładu 1. Każdy węzeł wewnętrzny jest etykietowany sumą prawdopodobieństw swoich synów.

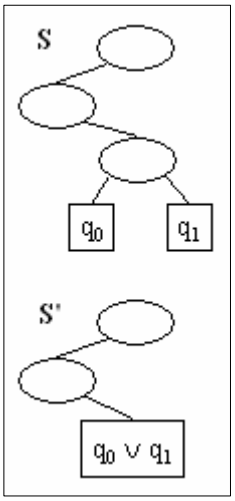



Teraz przedstawimy lemat, który posłuży do udowodnienia następnego twierdzenia.

**Lemat 1.** *Istnieje kod optymalny, w którym pewne dwa najdłuższe słowa mają tę samą długość, różnią się tylko ostatnim bitem (są braćmi) i odpowiadają symbolom o najmniejszych prawdopodobieństwach występowania (czyli dwóm najmniejszym wartościom ze zbioru p ).*

**Dowód.** Niech $c = \varphi_S(q)$ będzie najdłuższym słowem dla $q \in S$. Istnieje $c' = \varphi_S(q')$ takie, że $|c| = |c'|$, w przeciwnym przypadku moglibyśmy usunąć ostatni bit z $c$ i dostać kod, który jest nadal dekodowalny (nie psujemy warunku o bezprefiksowości), ale przeczyłoby to założonej optymalności (nowy kod byłby krótszy).
Jeśli $c$ nie miałby brata to znów moglibyśmy usunąć jego ostatni bit, zatem $c$ i $c'$ muszą być braćmi (różnią się tylko ostatnim bitem).
Ostatecznie $c$ i $c'$ muszą odpowiadać symbolom o najmniejszych prawdopodobieństwach występowania, gdyż tylko wtedy otrzymamy kod o najkrótszej średniej długości.

**Twierdzenie 1.** *Kod generowany przez algorytm Huffmana jest optymalnym bezprefiksowym kodem binarnym.*




**Dowód.** Udowodnimy optymalność poprzez indukcję ze względu na $|S|$.
Jeśli $|S| = 1$ to $L(\varphi^H_S) = 0$.
Przypuśćmy, że mamy $S$, $S' = (S - \{q_0, q_1\}) \cup \{q_0 \cup q_1\}$ jak w konstrukcji i $\varphi^H_{S'}$ optymalny. Popatrzmy na różnicę w długości tych kodów

$$L(\varphi^H_S) - L(\varphi^H_{S'}) = \underbrace{|\varphi^H_S(q_0)|}_{l} \cdot p(q_0) + \underbrace{|\varphi^H_S(q_1)|}_{l} \cdot p(q_1) - \underbrace{|\varphi^H_{S'}(q_0 \vee q_1)|}_{l-1} \cdot (p(q_0) + p(q_1)) =$$

$$= l \cdot p(q_0) + l \cdot p(q_1) - (l-1) \cdot (p(q_0) + p(q_1)) = p(q_0) + p(q_1)$$

Jak widać długość kodu w kolejnych krokach indukcyjnych różni się o pewną określoną wielkość niezależną od samego kodu.

Przyjmijmy, że $\psi$ to jakiś optymalny kod bezprefiksowy dla $S$. Wybieramy dwa najdłuższe słowa w kodzie $\psi$. Bez utraty ogólności możemy przyjąć, że to są właśnie $\psi(q_0)$ i $\psi(q_1)$ ($q_0$ i $q_1$ dobieraliśmy ze względu na prawdopodobieństwo, zatem im mniejsze prawdopodobieństwo tym dłuższy kod).
Moglibyśmy określić kod dla $S'$, w taki sposób, że $\psi'(q_0 \vee q_1) = \psi(q_0)\{0,1\}^{-1}$ (skracamy kod o jeden bit).
Wtedy, analogicznie do poprzednich obliczeń, zachodzi

$$L(\psi) - L(\psi') = |\psi(q_0)| \cdot p(q_0) + |\psi(q_1)| \cdot p(q_1) - |\psi'(q_0 \vee q_1)| \cdot (p(q_0) + p(q_1)) = \underbrace{p(q_0) + p(q_1)}_{\Delta}$$

Przypuśćmy teraz, że $\varphi_S^H$ jest kodem nie optymalnym, czyli zachodzi

$$L(\psi) < L(\varphi_S^H)$$

wtedy nowy kod dla $S'$

$$L(\psi') = L(\psi) - \Delta < L(\varphi_S^H) - \Delta = L(\varphi_{S'}^H)$$
$$L(\psi') < L(\varphi_{S'}^H)$$

co jest sprzeczne z założeniem o optymalności $\varphi_{S'}^H$. Zatem $\varphi_S^H$ musi być optymalny dla $S$.

## 3 Entropia przestrzeni produktowej

Blokowe kodowanie symboli polega na przypisywaniu słów kodowych o różnej liczbie bitów blokom o stałej długości $n$ symboli. Taka metoda umożliwia zmniejszenie średniej długości kodu przypadającego na symbol.

Przedstawimy jak obliczać **entropię przestrzeni produktowej** $\langle S^n, p \rangle$

Ogólniej: $\langle Z_1, p_1 \rangle$ $\langle Z_2, p_2 \rangle$

Określamy

$$\langle Z_1 \times Z_2, p \rangle, \; p(z_1, z_2) = p(z_1) \cdot p(z_2)$$

intuicyjnie (*)

$$H(Z_1 \times Z_2) = H(Z_1) + H(Z_2)$$

wynika to z liniowości wartości oczekiwanej zmiennej losowej.
Ogólnie

$$E(aX + bY) = aEX + bEY$$

Przypuśćmy, że mamy zmienne losowe $X$ na $Z_1$ i $Y$ na $Z_2$. Na $Z_1 \times Z_2$ określamy

$$(aX + bY)(Z_1, Z_2) = aX(Z_1) + bY(Z_2).$$

Rozważmy

$$X'(Z_1, Z_2) = X(Z_1) \text{ i } Y'(Z_1, Z_2) = Y(Z_2)$$

wtedy

$$EX = EX'$$
$$EY = EY'$$

zatem

$$E\left(aX' + bY'\right) = aEX + bEY$$

Zastosujemy to do naszego wzoru intuicyjnego (*)

$$H_k\left(Z_1 \times Z_2\right) = E\left(\lambda \cdot z_1 \cdot z_2 - \log_k p\left(z_1, z_2\right)\right) =$$
$$= E\left(\lambda \cdot z_1 \cdot z_2 - \log_k p\left(z_1\right) - \log_k p\left(z_2\right)\right) =$$
$$= E\left(\lambda \cdot z_1 - \log_k p\left(z_1\right)\right) + E\left(\lambda \cdot z_2 - \log_k p\left(z_2\right)\right) =$$
$$= H_k\left(Z_1\right) + H_k\left(Z_2\right)$$

## 4 Kod Shannona-Fano

Przypuśćmy najpierw, że $p(s) \neq 0$, dla $s \in S$. Określamy $l\left(s\right) = \left\lceil log_r\left(\frac{1}{p\left(s\right)}\right)\right\rceil$. Wtedy funkcja $l$ spełnia nierówność Krafta:

$$\sum_{s \in S} \frac{1}{r^{l\left(s\right)}} \leq 1$$

Istotnie $\frac{1}{r^{l\left(s\right)}} \leq \frac{1}{r^{\log\left(\frac{1}{p\left(s\right)}\right)}} = p\left(s\right)$

W kodzie bezprefiksowym nie zakodujemy słów o prawdopodobieństwie 0 (ponieważ nie będą one używane).

A zatem istnieje kod $\varphi$ o funkcji $\left|\varphi\right| = l$, wtedy średnia wartość długości takiego kodu wynosi

$$H_K\left(S\right) \leq L_K\left(\varphi\right) < \underbrace{\sum_{s \in S} p\left(s\right) \cdot \left(\log_r\left(\frac{1}{p\left(s\right)}\right) + 1\right)}_{H_K(s)+1}$$

Przypadek gdy istnieją $s$, takie że $p\left(s\right) = 0$.

Jeśli $\sum_{\substack{s \in S \\ p(s) \neq 0}} \frac{1}{r^{l\left(s\right)}} < 1$, to wtedy nie ma problemu – jest miejsce żeby zakodować pozostałe słowa (jest kod nie korzystający ze wszystkich liści w drzewie).

Jeśli $\sum_{\substack{s \in S \\ p(s) \neq 0}} \frac{1}{r^{l\left(s\right)}} = 1$ i powiedzmy, że $p\left(S_1\right) \neq 0$. Określmy nową funkcję:

$$l'(s_1) = l(s_1) + 1$$

$$l'(s_1) = l(s_1) \text{ dla } s \neq s_1$$

Wtedy „robi się miejsce”, przy czy nierówność $L_K(\varphi') \leq H_K(S) + 1$ (gdzie $\varphi'$ kod o $|\varphi'| = l'$) pozostaje zachowana (bo $\frac{1}{r^{l(s)}}$ jest $\leq p(s)$, sumując $\frac{1}{r^{l(s)}}$ powiększamy tylko jeden obiekt).

**Twierdzenie 1** *(Pierwsze twierdzenie Shannona)*

$$\frac{L_K(S^n)}{n} \to H_K(S),$$

dla $n \to \infty$, $S^n$ - przestrzeń produktowa n-tki obiektów z $S$ (z sumarycznym prawdopodobieństwem)

***Dowód.***

Mamy:

$$H_K(S^n) \leq L_K(S^n) \leq H_K(S^n) + 1$$

$$nH_K(S) \leq L_K(S^n) \leq nH_K(S) + 1 \ \| : n$$

$$H_K(S^n) \leq \frac{L_K(S^n)}{n} \leq H_K(S^n) + \frac{1}{n}$$

$$\frac{L_K(S^n)}{n} \to H_K(S)$$

□.

## 5 Entropia względna

Mamy kostkę o sześciu, ponumerowanych ścianach $\{1,2,3,4,5,6\}$, dla której prawdopodobieństwo wypadnięcia którejkolwiek z nich jest równe $p(i) = \frac{1}{6}$. Przypuśćmy, że ściany kostki są dodatkowo pokolorowane na trzy kolory, w taki sposób, że przeciwległe ściany mają te same kolory. Tak więc mamy zbiór kolorów $\{1,2,3\}$ dla którego prawdopodobieństwo wystąpienia jednego z kolorów wynosi $p(j) = \frac{1}{3}$.

Otrzymując dwie funkcje nie możemy „tak po prostu” policzyć całkowitego prawdopodobieństwa przez mnożenie $p(i)$ z $p(j)$ (ponieważ jak wiadomo niektóre z sytuacji nie będą mogły zajść – niektóre z kolorów nigdy nie wystąpią z konkretną liczbą).

Mamy dwie zmienne losowe:

$$\text{Ściany} \rightarrow \{1,2,3,4,5,6\} \text{ oraz}$$

$$\text{Ściany} \rightarrow \{1,2,3\}$$

Skończona przestrzeń probabilistyczna $(\Omega, p)$, $\sum_{\omega\in\Omega} p(\omega) = 1$

Dla $z \subseteq \Omega$, $p(z) = \sum_{\omega\in z} p(\omega)$

*Zmienna losowa* $X : \Omega \rightarrow$ *Zbiór*

$p(X = z) = \sum_{\omega : X(\omega)=z} p(\omega)$ (czasem będziemy zapisywać $p(z)$ zamiast $p(X = z)$ )

*Wartość oczekiwana* $E(X) = \sum_{\omega\in\Omega} p(\omega)\cdot X(\omega)$

***Uwaga*** *(odnośnie notacji)* Dla funkcji $X$, zbiorem wartości jest $\overrightarrow{X}$. Jeśli $A : \Omega \rightarrow \{a_1,...,a_m\}$ jest zmienną losową, to

$$H_K(A) \underset{def.}{=} \sum_{i=1}^{m} p(A = a_i)\left(-\log_K p(A = a_i)\right) = E\left(\lambda\omega \in \Omega. - \log_K p(A = A(\omega))\right)$$

*Entropia względna* „inaczej” to:

- średnia ilość pytań w grze przy optymalnej strategii
- minimalna średnia długość kodu

## 6 Entropia warunkowa

Załóżmy, że $A : \Omega \rightarrow \{a_1,...,a_m\}$ oraz $B : \Omega \rightarrow \{b_1,...,b_n\}$ to zmienne losowe

$H_K(A|b_j) \underset{def.}{=} \sum_{i=1}^{m} p(a_i|b_j)\cdot \log_K\left(\frac{1}{p(a_i|b_j)}\right)$, gdzie $b_j$ jest podpowiedzią.

*Entopią warunkową* będzie $H_K(A|B) = \sum_{j=1}^{n} p(b_j)\cdot H(A|b_j)$

Możemy także utworzyć nową zmienna $AB:\omega\to\langle A(\omega),B(\omega)\rangle$, gdzie $\omega$ pochodzi z jednego rzutu. Wtedy

$$H_K(AB)=\sum_{i=1}^{m}\sum_{j=1}^{n}p(a_i\cap b_j)\cdot\log_K\left(\frac{1}{p(a_i\cap b_j)}\right)$$

*Przypomnijmy, że* $A$ *i* $B$ *są niezależne jeśli* $p(A=a_i\wedge B=b_j)=p(A=a_i)\cdot p(B=b_j)$, *albo w skrócie* $p(a_i\cap b_j)=p(a_i)\cdot p(b_j)$

Jeśli $A$ i $B$ są niezależne, to

$$H_K(AB)=\sum_{i=1}^{m}\sum_{j=1}^{n}p(a_i)\cdot p(b_j)\cdot\left(\log_K\frac{1}{p(a_i)}\cdot\log_K\frac{1}{p(b_j)}\right)=H_K(A)+H_K(B)$$

W ogólnym przypadku (gdy $A$ i $B$ mogą być zależne), mamy „tylko”:

$$H_K(AB)=\sum_{i=1}^{m}\sum_{j=1}^{n}p(b_j|a_i)\cdot p(a_i)\cdot\left(\log_K\frac{1}{p(b_j|a_i)}\cdot\log_K\frac{1}{p(a_i)}\right)=$$

$$=\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}p(b_j|a_i)\cdot p(a_i)\cdot\log_K\frac{1}{p(b_j|a_i)}}_{H_K(B|A)}+\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}\underbrace{p(b_j|a_i)}_{b_j=1\text{ dla ustalonego j}}\cdot p(a_i)\cdot\log_K\frac{1}{p(a_i)}}_{H_K(A)}=$$

$$=H_K(B|A)+H_K(A)$$

**Fakt.** $H_K(AB)=H_K(A)+H_K(B|A)=H_K(B)+H_K(A|B)$

$$H_K(A)+H_K(B)=\sum_{i=1}^{m}p(a_i)\cdot\log_K\left(\frac{1}{p(a_i)}\right)+\sum_{j=1}^{n}p(b_j)\cdot\log_K\left(\frac{1}{p(b_j)}\right)=$$

$$=\left(\sum_{i=1}^{m}\sum_{j=1}^{n}p(a_i\cap b_j)\cdot\log_K\left(\frac{1}{p(a_i)}\right)\right)+\left(\sum_{i=1}^{m}\sum_{j=1}^{n}p(a_i\cap b_j)\cdot\log_K\left(\frac{1}{p(b_j)}\right)\right)=$$

$$=\sum_{i=1}^{m}\sum_{j=1}^{n}p(a_i\cap b_j)\cdot\left(\log_K\frac{1}{p(a_i)\cdot p(b_j)}\right)=$$

= **sytuacja: ZŁOTY LEMAT ( czyli:** $\sum p_i\log\frac{1}{p_i}\le\sum p_i\log\frac{1}{q_i}$ **)**

Oczywiście należy wytłumaczyć skąd się wzięły powyższe przekształcenia – otóż korzystaliśmy z:

1. $p(a_i) = \sum_{j=1}^{n} p(a_i \cap b_j)$
2. $\sum_{i,j} p(a_i) \cdot p(b_j) = 1$
3. Jeśli $p(a_i) \cdot p(b_j) = 0$, to $p(a_i \cap b_j) = 0$

*Z powyższych wynika* **Fakt.**

$H_K(A|B) \le H_K(A)$ i symetrycznie $H_K(B|A) \le H_K(B)$, przy czym równość występuje dokładnie wtedy, gdy $A$ i $B$ są niezależne.

*Wyjaśnienie faktu* – wiemy więcej: wiemy że B i jakieś A, więc entropia jest mniejsza.

**Uwaga.**

$$H_K(A) - H_K(A|B) = H_K(B) - H_K(B|A), \text{ ponieważ}$$

powyższe są równe $H_K(A) + H_K(B) - H_K(AB)$

Przedstawione powyżej różnice ($H_K(A) - H_K(A|B) = H_K(B) - H_K(B|A)$) oznacza się również jako $I(A,B)$ lub $I(B,A)$ i nazywa ***wzajemną informacją*** A o B (lub B o A)

## 7 Konkluzje dotyczące entropii względnej

$$H_K(A) + H_K(B) \ge H_K(AB) = H_K(A|B) + H_K(B) = H_K(B|A) + H_K(A)$$

całkowita równość zachodzi jedynie wtedy, gdy $A$ i $B$ są niezależne.

$$H_K(A) - H_K(A|B) = H_K(B) - H_K(B|A) = I(A,B),$$

gdzie $I(A,B)$ to wzajemna informacja między $A$ i $B$.

$$H_K(A|B) = \sum_b \underbrace{\left( \sum_a p(a|b) \cdot \log\left( \frac{1}{p(a|b)} \right) \right)}_{H_K(A|b) \text{ (przy konwencji, że B=b)}} \cdot p(b)$$

$$H_K\left(A\middle|B\right) \le H_K\left(A\right)$$

*Entropia* jest miarą trudności (złożoności) obiektów

*Przykład.*

Załóżmy, że dwie osoby X i Y chcą się spotkać. Mamy dwie zmienne losowe: A, która opisuje położenie Xa i B, która opisuje czas (w sensie godziny).

W średnim przypadku Y ma większe szanse na spotkanie z X jeśli zna czas (w którym może znaleźć X w określonym miejscu):

$$H_K\left(A\middle|B\right) \le H_K\left(A\right)$$

**Uwaga 1**. Jednak istnieją pewne określone wartości $B = b$ (takie momenty czy też przedziały czasu), w których znalezienie Xa jest trudniejsze niż średnio:

$$H_K\left(A\middle|b\right) > H_K\left(A\right)$$

**Uwaga 2.** Nie należy mylić tego doświadczenia z doświadczeniem, w których występują dwa eksperymenty – jeden z dwoma wartościami (kolory oraz liczby na kostce).