Izvorni znanstveni članak

# Korelacija, informacija i kauzalnost

Božidar Tepeš
Jacinta Grbavac
Odsjek za informacijske znanosti, Filozofski fakultet
Ivana Lučića 3, Zagreb, Hrvatska
btepes@ffzg.hr

**Sažetak**

*Statistička povezanost zove se korelacija, a stupanj statističke povezanosti mjeri se koeficijentom korelacije. U prirodi jedan od glavnih zakonitosti je kauzalnost, a kauzalnost ne vidimo iz nezavisnosti i povezanosti. Kauzalnost pretpostavlja relaciju između slučajne varijable koja je uzrok i slučajne varijable koja je posljedica. U slučaju nezavisnosti relacija između slučajnih varijabli je komutativna ili govorimo o skupu nezavisnih slučajnih varijabli, a ne o nizovima slučajnih varijabli kakve imamo kod kauzalnosti. Kauzalnost se prikazuje acikličkim usmjerenim grafom koji predstavlja kauzalnu povezanost slučajnih varijabli. Informacija u radu je mjera stupnja sličnosti između statističkog modela odnosa slučajnih varijabli i stvarnih odnosa varijabli mjerenja dobivenih uzorkom. U radu će se pojmovi korelativnosti, informacije i kauzalnosti ilustrirati Markovljevim i kauzalnim modelom predikata hrvatskog jezika*

**Ključne riječi**: Korelacija, informacija, kauzalnost, Markovljev model, kauzalni model, predikat hrvatskog jezika.

## 1. Korelacija

Koeficijent korelacije [1] govori o nezavisnosti pojava. Nezavisnost je pojam vezan za vjerojatnost i slučajne varijable, a statistika koristi uzorak koji je niz slučajnih varijabli koje imaju svoje razdiobe. Ako nezavisnost slučajne varijable $X$ i slučajne varijable $Y$ označimo $X \perp Y$ tada za njihove funkcije razdiobe vrijedi:

$$f(x,y) = f(x) \cdot f(y) \tag{1.1}$$

Koeficijent korelacije slučajnih varijabli $X$ i $Y$ označimo s $r_{XY}$ . . Slučajne varijable mogu biti nezavisne i tada je koeficijent korelacije jednak nuli. Obrat ne vrijedi ili ako je koeficijent korelacije jednak nuli slučajne varijable ne moraju biti nezavisne. Tu činjenicu možemo zapisati relacijom (1.2).

(1.2) $$\left(X \perp Y\right) \Rightarrow \left(r_{XY} = 0\right)$$

U slučaju niza slučajnih varijabli $X_1, X_2, ..., X_n$ možemo govoriti o uvjetnoj nezavisnosti dvije slučajne varijable $X_i$ i $X_j$ iz niza slučajnih varijabli $X_1, X_2, ..., X_n$ uz uvjet ostalih slučajnih varijabli iz niza koje se razlikuju od slučajnih varijabli $X_i$ i $X_j$. Uvjetnu nezavisnost označimo $X_i \perp X_j \mid X_1,...,X_{i-1},X_{i+1},...,X_{j-1},...,X_{j+1},...,X_n$ i za uvjetnu nezavisnost vrijedi jednadžba (1.3).

(1.3)
$$f\left(x_1,x_2,...,x_n\right) = f\left(x_i \mid x_1,...,x_{i-1},x_{i+1},...,x_n\right) \cdot f\left(x_j \mid x_1,...,x_{j-1},x_{j+1},...,x_n\right)$$

Uvjetna nezavisnost mjeri se koeficijentom parcijalne korelacije $\rho_{X_iX_j \cdot Z}$ gdje smo sa $Z$ označili skup slučajnih varijabli koje se razlikuju od $X_i$ i $X_j$ ili $Z = \left\{X_1,...,X_{i-1},X_{i+1},...,X_{j-1},...,X_{j+1},...,X_n\right\}$. Također vrijedi relacija (1.4) između vrijednosti koeficijenta parcijalne korelacije i uvjetne nezavisnosti.

(1.4) $$\left(X_i \perp X_j \mid Z\right) \Rightarrow \left(\rho_{X_iX_j \cdot Z} = 0\right)$$

Obrat relacija (1.2) i (1.4) vrijedi u slučaju normalne razdiobe slučajnih varijabli. To znači ako su koeficijent korelacije i koeficijent parcijalne korelacije jednaki nuli onda su slučajne varijable nezavisne i uvjetno nezavisne. To znači da možemo govoriti o nezavisnosti za male vrijednosti koeficijenata korelacije.

## 2. Informacija

Pojam informacije temelji se na pojmu entropije ili neodređenosti. Možemo reći informacija je razlika entropija ili smanjenje neodređenosti. U slučaju jedne diskretne slučajne varijable $X$ koja ima vrijednosti $x$ s vjerojatnostima $f\left(x\right)$ za koje vrijedi $\sum_x f\left(x\right) = 1$ entropija $H\left(X\right)$ je određena:

(2.1) $$H\left(X\right) = -\sum_x f\left(x\right) lb\left(f\left(x\right)\right)$$

gdje je $lb\left(f\left(x\right)\right) = \log_2\left(f\left(x\right)\right)$ logaritam po bazi dva. Kažemo slučajna varijabla $X$ ima entropiju $H\left(X\right)$ bita.

Statističkim modelima na temelju uzorka ocjenjujemo razdiobu slučajne varijable $X$. Ako na temelju modela $M$ ocijenimo vjerojatnosti slučajne varijable $X$ s $m(x)$. Temeljno pitanje je koliko naš model odgovara stvarnim vrijednostima ili koja je razlika između $f(x)$ i $m(x)$. Ocjenu neodređenosti modela daje nam ukrštena entropija $H(X,M)$:

$$H(X,M) = -\sum_x f(x) lb(m(x)) \tag{2.2}$$

Informacija o točnosti modela je razlika između ukrštene entropije $H(X,M)$ i entropije $H(X)$. Ta informacija zove se relativna entropija ili Kullback-Leiblerova razlika [2] $I(f \parallel m)$:

$$\begin{aligned} I(f \parallel m) &= H(X,M) - H(X) \\ &= \sum_x f(x) lb \tfrac{f(x)}{m(x)} \end{aligned} \tag{2.3}$$

Može se pokazati nenegativnost Kullback-Leiblerove razlike ili $I(f \parallel m) \geq 0$. KL razlika je informacija dobivena uvidom u stvarne vrijednost slučajne varijable ili neodređenost modela koja se smanjila uvidom u stvarne podatke o slučajnoj varijabli. Informacija kao razlika entropija također se mjeri u bitima. Minimalizacijom Kullback-Leiblerove razlike možemo ocijeniti model koji se najmanje razlikuje od stvarnih podataka.

Markovljev model $M$ je stohastički model koji preko Markovljevog lanca opisuje razdiobu $m(x)$ slučajne varijable $X$. Markovljev lanac prvog reda određuje razdiobu $m(x_i)$ stanja $i$ na temelju razdiobe $m(x_{i-1})$ stanja $i-1$ koje prethodi stanju $i$. Povezanost je određena vjerojatnostima prijelaza $p_{i-1,i}$:

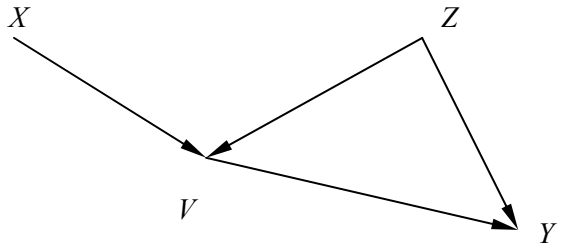
$$m(x_i) = \sum_{i-1} p_{i-1,i} m(x_{i-1}) \tag{2.4}$$

Markovljev lanac može biti i višeg reda $r$ govorimo o $r-$ gramskim modelima u kojima su vjerojatnosti prijelaza iz $r$ prethodnih stanja $i-1, i-2, ..., i-r$ u $i-$to stanje koje slijedi iza prethodnih stanja.

## 3. Kauzalnost

Kauzalna povezanost varijabli mjerenja određena je smjerom od varijable koja je uzrok prema varijabli koja je posljedica. Tako se grafički kauzalne poveza-

nosti prikazuju acikličkim usmjerenim grafom koji se naziva kauzalna struktura. Kauzalna povezanost može biti direktna ako postoji usmjereni put između dvije varijable ili zbunjujuća kauzalna povezanost ako postoji put koji nije usmjeren između dvije varijable. Primjere direktne i zbunjujuće kauzalnosti prikazane su na kauzalnoj strukturi na slici 1. gdje je direktna kauzalnost između uzroka $X$ i posljedice $Y$ preko varijable $V$ i zbunjujuća kauzalnost između uzroka $X$ i posljedice $Y$ preko varijable $Z$.




Slika 1. Kauzalna struktura

Pomoću IC* algoritma induktivne kauzalnosti [3], [4] određuje se kauzalna struktura ili aciklički usmjereni graf koji pokazuje kauzalnu povezanost varijabli mjerenja. Intenzitet kauzalne zavisnosti mjeri se informacijom o međuzavisnosti [5]:

$$I(X \perp Y \mid Z) = -0{,}5\, lb\left(1 - \rho^2_{XY \cdot Z}\right) \tag{3.1}$$

gdje je $lb$ logaritam po bazi 2, a $Z$ je podskup skupa varijabli mjerenja u kojem nisu varijable $X$ i $Y$.
Kauzalni model ili kauzalna mreža je kauzalna struktura na kojoj varijablama mjerenja ili vrhovima $v$ su pridružene funkcije $f_v$ oblika:

$$v = f_v\left(R, u_v\right) \tag{3.2}$$

gdje su $R$ roditelji vrha $v$, a $u_v$ su nezavisne slučajne varijable. Funkcije $f_v$ mogu biti razdiobe slučajne varijable $v$ ili $p(v)$koje se mogu odrediti iz uvjetnih razdioba $p(v \mid R)$ za koje vrijedi:

$$p(v) = p(v \mid R) \cdot p(R) \tag{3.3}$$

Jednadžba se rekurzivno primjenjuje na cijeloj kauzalnoj strukturi i određuje razdiobu kauzalnog modela $p(km)$:

$$p(km) = \prod_{v,R} p(v \mid R) \tag{3.4}$$

Informacija o točnosti kauzalnog modela je informacija o točnosti statističkoga modela određena Kullback-Leiblerova razlika $I(f \,||\, m)$ određena izrazom (2.3) gdje je $f$ stvarna razdioba $p(v)$ varijabli mjerenja, a $m$ je razdioba $p(km)$ određena kauzalnim modelom.

## 4. Model predikata hrvatskog jezika

Pojam korelacije, informacije i kauzalnosti ćemo ilustrirati na primjeru predikata hrvatskog jezika. Iz baze označenih rečenica hrvatskog jezika [6] uzet je uzorak od stotinu rečenica. U tim rečenicama označeni su dijelovi predikata: pomoćni glagol *I*, glavni glagol *V*, odrednica (zamjenica) *D*, pridjev *A* i imenica *N*. Koeficijenti korelacije $r_{ij}$ su prikazani matricom $R$:

$$R = \begin{bmatrix} 1 & 0{,}99 & 0{,}97 & 0{,}96 & 0{,}94 \\ 0{,}99 & 1 & 0{,}97 & 0{,}96 & 0{,}94 \\ 0{,}97 & 0{,}97 & 1 & 0{,}94 & 0{,}95 \\ 0{,}96 & 0{,}96 & 0{,}94 & 1 & 0{,}95 \\ 0{,}94 & 0{,}94 & 0{,}95 & 0{,}95 & 1 \end{bmatrix} \tag{4.1}$$

Iz matrice koeficijenata korelacije $R$ vidimo jaku povezanost pojedinih dijelova predikata. Parcijalni koeficijenti korelacije $\rho_{ij \cdot Z}$ između dijelova predikata $i$ i $j$ uz uvjet skupa $Z$ u kojem nisu $i$ i $j$ prikazani su matricom $\rho$:

$$\rho = \begin{bmatrix} - & 0{,}85 & -0{,}54 & 0{,}16 & 0{,}17 \\ 0{,}85 & - & 0{,}44 & 0{,}13 & -0{,}21 \\ -0{,}54 & 0{,}44 & - & -0{,}16 & 0{,}44 \\ 0{,}16 & 0{,}13 & -0{,}16 & - & 0{,}49 \\ 0{,}17 & -0{,}21 & 0{,}44 & 0{,}49 & - \end{bmatrix} \tag{4.2}$$

Parcijalne koeficijente korelacije možemo podijeliti na dvije skupine. U prvoj skupini su koeficijenti veći od 0,4, a u drugoj su skupini koeficijenti manji od 0,4. Prva skupina koeficijenata pokazuje povezanost, a druga skupina nezavisnost dijelova predikata.
Iz uzorka od stotinu rečenica hrvatskog jezika izračunata je razdioba dijelova predikata $f(x)$:

| $x$ | $I$ | $V$ | $D$ | $A$ | $N$ |
|---|---|---|---|---|---|
| $f(x)$ | 0,30 | 0,49 | 0,08 | 0,07 | 0,06 |

Tabela 4.1 Razdioba dijelova predikata na uzorku

Uz pretpostavku uniformne razdiobe dijelova predikata entropija je 2,81 bit, a entropija uzorka je 2,07 bita. Informacija o razdiobi uzorka je razlika entropija i iznosi 0,74 bita.
Sada pristupimo modeliranju predikata Markovljevim modelom (MM) i kauzalnim modelom (KM).
U Markovljevom modelu predikata bitna je matrica prijelaza $P$ s elementima $p_{ij}$ koji pokazuju vjerojatnosti nalaženja dijela predikata $j$ iza dijela predikata $i$:

$$P = \begin{bmatrix} 0{,}021 & 0{,}575 & 0{,}043 & 0{,}255 & 0{,}106 \\ 0{,}568 & 0{,}189 & 0{,}216 & 0{,}027 & 0{,}000 \\ 0{,}000 & 0{,}778 & 0{,}000 & 0{,}111 & 0{,}111 \\ 0{,}100 & 0{,}200 & 0{,}000 & 0{,}200 & 0{,}500 \\ 0{,}200 & 0{,}600 & 0{,}200 & 0{,}000 & 0{,}000 \end{bmatrix} \tag{4.3}$$

Uz pretpostavku početnog stanja uniformne razdiobe dijelova predikata i sustavnog množenja tog stanja s matricom prijelaza $P$ dobivamo:

(4.4)

$$\begin{bmatrix} 0{,}2 & 0{,}2 & 0{,}2 & 0{,}2 & 0{,}2 \end{bmatrix} \cdot \begin{bmatrix} 0{,}021 & 0{,}575 & 0{,}043 & 0{,}255 & 0{,}106 \\ 0{,}568 & 0{,}189 & 0{,}216 & 0{,}027 & 0{,}000 \\ 0{,}000 & 0{,}778 & 0{,}000 & 0{,}111 & 0{,}111 \\ 0{,}100 & 0{,}200 & 0{,}000 & 0{,}200 & 0{,}500 \\ 0{,}200 & 0{,}600 & 0{,}200 & 0{,}000 & 0{,}000 \end{bmatrix}^{n}$$

$$\underset{n\to\infty}{\longrightarrow} \begin{bmatrix} 0{,}27 & 0{,}40 & 0{,}12 & 0{,}11 & 0{,}10 \end{bmatrix}$$

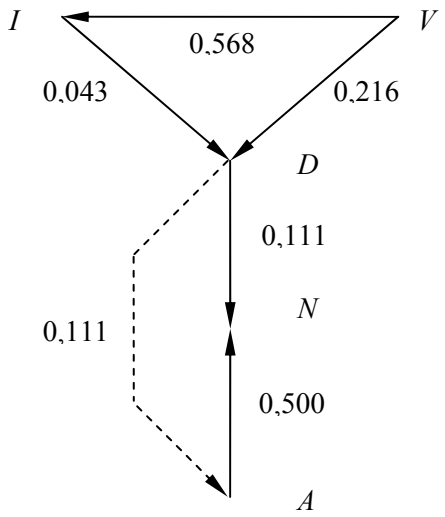
Iz toga slijedi razdioba dijelova predikata na temelju Markovljevog modela:

$$m_{MM} = \begin{bmatrix} 0{,}27 & 0{,}40 & 0{,}12 & 0{,}11 & 0{,}10 \end{bmatrix} \tag{4.5}$$

Kullback-Leiblerova razlika je:

$$\begin{aligned} I(f \,||\, m_{MM}) &= 0{,}30 lb \tfrac{0{,}30}{0{,}27} + 0{,}49 lb \tfrac{0{,}49}{0{,}40} + 0{,}08 lb \tfrac{0{,}08}{0{,}12} + 0{,}07 lb \tfrac{0{,}07}{0{,}11} \\ &+ 0{,}06 lb \tfrac{0{,}06}{0{,}10} = 0{,}052\, bita \end{aligned} \tag{4.6}$$

i pokazuje razliku između Markovljevog modela i uzorka predikata hrvatskog jezika.
Kod kauzalnog modela koristimo kauzalnu strukturu dobivenu iz parcijalnih koeficijenata korelacije:




Slika 4.1 Kauzalna struktura predikata

Iz kauzalne strukture je vidljiva dominantna uloga glagola $V$ koji je uzrok nalaženja svih ostalih dijelova predikata jer glagol otvara u predikatu mjesta svim ostalim dijelovima predikata $V$, $D$, $A$ i $N$. To je u suglasnosti s gramatičkim ustrojstvom rečenica hrvatskog jezika gdje je glagol temeljna riječ predikata [7].
Iz direktnih kauzalnih i zbunjujućih kauzalnosti dobivamo razdiobu dijelova predikata kako slijedi:

$$m_{KM}(I) = p_{21} = 0{,}568 \approx 0{,}57$$
$$m_{KM}(D) = p_{23} + p_{21}p_{23} = 0{,}216 + 0{,}568 \cdot 0{,}043 = 0{,}240 \approx 0{,}24$$
$$m_{KM}(A) = p_{34} = 0{,}111 \approx 0{,}11$$
$$m_{KM}(N) = (p_{23} + p_{21}p_{23})p_{35} + p_{34}p_{45} = 0{,}240 \cdot 0{,}111 + 0{,}111 \cdot 0{,}500$$
$$= 0{,}082 \approx 0{,}08$$

Razdioba dijelova predikata $I$, $D$, $A$ i $N$ izvedena iz razdiobe svih dijelova predikata $I$, $V$, $D$, $A$ i $N$ na uzorku od stotinu rečenica hrvatskog jezika je:

| $x$ | $I$ | $D$ | $A$ | $N$ |
|---|---|---|---|---|
| $f_1(x)$ | 0,58 | 0,17 | 0,14 | 0,11 |

Tabela 4.2 Razdioba dijelova predikata iz uzorka bez $V$

Kullback-Leiblerova razlika je:

$$I(f_1 \,||\, m_{KM}) = 0{,}58 lb \tfrac{0{,}58}{0{,}57} + 0{,}17 lb \tfrac{0{,}17}{0{,}24} + 0{,}14 lb \tfrac{0{,}14}{0{,}11} + 0{,}11 lb \tfrac{0{,}11}{0{,}08} = 0{,}029\, bita \tag{4.7}$$

i pokazuje razliku između Markovljevog modela i uzorka predikata hrvatskog jezika. Dobiveni rezultat pokazuje prednost kauzalnog modela pred Markovljevim jer je Kullback-Leiblerova razlika kauzalnog modela manja jer je razlika modela od uzorka manja. Cijeli model bi trebalo testirati na cijeloj Baza morfološki i sintaktički označenih rečenica hrvatskog jezika i proširiti na ostale dijelove rečenica hrvatskog jezika i tako definirati kauzalnu strukturu cijelih rečenica.

## Literatura

[1] I. Pavlić, Statistička teorija i primjena (1988), Tehnička knjiga, Zagreb
[2] V. Hatzivassiloglou (2006), *Cross Entropy and Relative Entropy,* University of Texas,Dallas, http://www.hlt.utdallas.edu/~vh/Courses/Spring06/StatisticalNLP.html
[3] J. Pearl (2001), *Causality: Models, Reasoning, and Inference,* Cambridge University Press, Cambridge UK
[4] B. Tepeš (2007), *Statistički modeli na grafovima,* (skripte) Filozofski fakultet, Zagreb, http://www.ffzg.hr
[5] J. Whittaker (1990), *Graphical Models in Applied Multivariate Statistics,* John Wiley & Sons, Chichester UK
[6] Baza morfološki i sintaktički označenih rečenica hrvatskog jezika (2006), znanstveni projekt Označavanje i prepoznavanje riječi hrvatskog jezika, voditelj B. Tepeš, http://infoz.ffzg.hr/tepes/
[7] R. Katičić (1991), *Sintaksa hrvatskog književnog jezika*, HAZU, Globus, Zagreb