INFORMAČNÍ OBSAH ZNALOSTÍ

Arnošt Veselý
Česká zemědělská univerzita, Kamýcká , Praha 6 – Suchdol, vesely@pef.czu.cz

ABSTRAKT:
Příspěvek se zabývá problémem hodnocení a kvantitativního vyjádření „informativnosti“ znalostí. Je popsána metoda, založená na principech teorie informace, která umožňuje „informativnost“ znalostí kvantifikovat.

KLÍČOVÁ SLOVA:
znalosti, dobývání znalostí, informační obsah znalostí

Získávání dříve neznámých a potenciálně užitečných znalostí, které jsou implicitně obsaženy v souborech dat či v databázích je důležité, protože získání nových znalostí umožňuje kvalitnější rozhodování. V oblasti nazývané dobývání dat (data mining) byla navržena řada metod jak znalosti ze souborů dat automaticky získat. Vzniká proto otázka, jak tyto znalosti navzájem porovnávat a hodnotit z hlediska jejich „informativnosti“. Příspěvek se zabývá problémem hodnocení a kvantitativního vyjádření „informativnosti“ znalostí.

Pojem znalost souvisí s modelem jazyk – ontologie a v rámci tohoto modelu lze pojem znalost přesně vymezit. Tento model předpokládá, že existuje svět (ontologie) skládající se z objektů, které mají určité vlastnosti a mezi kterými platí určité vztahy (relace). Dále se předpokládá, že pro popis ontologie je k dispozici jazyk, obvykle predikátový počet nebo jeho fragment. Pro objekty ontologie a jejich vlastnosti existují v jazyce jména. Objekty se označují individuálními konstantami a vlastnosti objektů jednomístnými predikáty. Jazyk dále obsahuje logické spojky non ($\neg$), and ($\wedge$), or $\vee$, imply ($\supset$) atd., které dovolují z jednoduchých výroků vytvářet složené výroky. Např. tvrzení, že objekt a má vlastnost P a Q se vyjádří jako $P(a) \wedge Q(a)$ atd.

Ontologie může být v různých stavech. Předpokládejme, že jazyk J popisující ontologii Ω obsahuje predikáty $\pi_1 , \ldots , \pi_r$, které označují vlastnosti objektů ontologie Ω a individuální konstanty $\omega_1 , \ldots , \omega_s$, které označují jednotlivé objekty. Každý stav ontologie Ω jednoznačně určuje pro každé j=1, . . r a pro každé k=1, . . . , s , zda objekt ω_j má vlastnost π_k nebo nemá. Proto každý stav ontologie Ω lze jednoznačně popsat konjunkcí α

$$\alpha = \alpha(\omega_1) \wedge \ldots \wedge \alpha(\omega_j) \wedge \ldots \wedge \alpha(\omega_s), \text{ kde}$$
$$\alpha(\omega_j) = l_1(\omega_j) \wedge \ldots \wedge l_i(\omega_j) \wedge \ldots \wedge l_r(\omega_j) \text{ a} \qquad (1)$$
$$l_i(\omega_j) = \pi_i(\omega_j), \text{ jestliže } \omega_j \text{ má vlastnost } \pi_i ,$$
$$l_i(\omega_j) = \neg\pi_i(\omega_j), \text{ jestliže } \pi_i \text{ nemá vlastnost } \pi_i .$$

Formule α, které jednoznačně popisují stavy ontologie, budeme nazývat atomy. V predikátovém počtu platí, že každou formuli β jazyka J popisujícího konečnou ontologii Ω, lze ekvivalentně vyjádřit jako disjunkci atomů (viz např. Mendelsson, 1997), tj. jako

$$\beta \equiv \vee_{i \in I_\beta} \alpha_i , \quad \text{kde } I_\beta \subset N \cup \{0\} \text{ a } N \text{ je množina přirozených čísel.}$$

Zde I_β je množina indexů těch atomů, které popisují stavy ontologie, ve kterých je formule β splněna.

V rámci výše uvedeného jednoduchého sémantického modelu jazyka budeme znalosti definovat jako formule, které jsou v ontologii, kterou jazyk popisuje, pravdivé. Jakmile jsme pojem znalost přesně vymezili lze se zamýšlet nad tím, jak jednotlivé znalosti navzájem porovnávat. Je zřejmé, že některé znalosti poskytují o dané ontologii zajímavější a důležitější výpovědi než jiné. Například tautologie (tj. formule platné vždy a tedy na ontologii nezávislé) žádnou informaci o ontologii neposkytují.

Standardním způsobem jak porovnávat znalosti je použít relaci logického vyplývání. Pokud β_1 a β_2 jsou dvě formule a β_2 logicky vyplývá z β_1 (tj. pokud $\beta_1 \supset \beta_2$ je tautologie) řekneme, že, β_1 je informativnější než β_2 . Relace logického vyplývání je reflexivní a transitivní a proto definuje na množině znalostí částečné uspořádání.

Výše uvedeným způsobem lze však hodnotit „informativnost“ znalostí jen velmi omezeným způsobem, protože:

a) porovnávat mezi sebou můžeme pouze některé znalosti,

b) porovnání je pouze kvalitativní a nedovoluje kvantitativní vyjádření.

V dalším ukážeme jak lze zavést kvantitativní míru „informativnosti“ znalostí, která je založena na principech teorie informace. Tento přístup k měření informačního obsahu znalostí byl poprvé navržen a publikován v práci (Vajda, Veselý, Zvárová, 2005).

Informace jako míra poklesu neurčitosti našeho obrazu světa po přijetí určité zprávy byla zavedena Shannonem a našla velké množství důležitých praktických aplikací. Základní představa je následující. Stavy ontologie se považují za realizace náhodné proměnné ξ. Pokud známe pravdivostní distribuci ξ, můžeme stanovit entropii $H(\xi)$, která vyjadřuje průměrnou míru neurčitosti s jakou lze odhadnout, který ze stavů ontologie nastane nebo nastal. Dále se předpokládá, že je k dispozici zpráva charakterizovaná náhodnou proměnnou η. Pokud známe podmíněné rozložení pravděpodobnosti $\xi \mid \eta$, lze stanovit podmíněnou entropii $H(\xi \mid \eta)$. Podmíněná entropie má hodnotu 0 pokud realizace zprávy jednoznačně určuje stav ontologie a hodnotu $H(\xi)$ pokud jsou náhodné proměnné ξ a η nezávislé. Informace, která je obsažena ve zprávě se pak definuje

$$I(\xi ; \eta) = H(\xi) - H(\xi \mid \eta)$$

a vyjadřuje míru snížení neurčitosti našeho obrazu světa po přijetí zprávy.

Předpokládejme, že $\beta_1,\ldots,\beta_m$ jsou znalosti a že současná znalost o pravděpodobnostním rozložení stavů ontologie je $q = \{q_i\}$.

Informaci $I(\beta_1,\ldots,\beta_m)$ obsaženou v znalostech $\beta_1,\ldots,\beta_m$ definujeme následovně:

$$I(\beta_1,\ldots,\beta_m) = -\log \sum_{i \in I_{\beta 1} \cap \ldots \cap I_{\beta m}} q_i \ . \tag{2}$$

Zaveďme následující φ-transformaci $\tilde{q} = \varphi(q; \beta_1,\ldots,\beta_m)$:

$$\tilde{q}_i = 0 \qquad \text{if } i \notin I_{\beta 1} \cap \ldots \cap I_{\beta m} , \tag{3}$$

$$\tilde{q}_i = \frac{q_i}{\sum_{i \in I_{\beta 1} \cap \ldots \cap I_{\beta m}} q_i} \qquad \text{if } i \in I_{\beta 1} \cap \ldots \cap I_{\beta m} . \tag{4}$$

Platí následující věta (důkaz viz (Vajda, Veselý, Zvárová, 2005)).

Věta Informace $I(\beta_1,\ldots,\beta_m)$ obsažená v znalostech $\beta_1,\ldots,\beta_m$ má následující vlastnosti:

1. $I(\beta_1,\ldots,\beta_m) \geq 0$.

2. Pokud z platnosti β_j logicky vyplývá platnost β_k, potom $I(\beta_j) \geq I(\beta_k)$.

3. Je-li $I(\beta_1,\ldots,\beta_m) > I(\gamma_1,\ldots,\gamma_m)$, potom

$$D(p||\varphi(q;\beta_1,\ldots,\beta_m) < D(p||\varphi(q;\gamma_1,\ldots,\gamma_m)),$$

kde p je skutečné pravděpodobnostní rozdělení stavů ontologie a

$$D(p||q) = \sum_i p_i \log\frac{p_i}{q_i}, \qquad (5)$$

$p\log p/q=0$ if $p=0$, $q\geq 0$ a $p\log p/q = \infty$ if $p>0$, $q=0$ je Leiber-Kullbackova divergence distribucí p a q (viz např. Cover T.M., 2006).

Definovali jsme míru množství informace obsažené ve znalostech $\beta_1,\ldots,\beta_m$. Vlastnost této míry lze jednoduše popsat takto: Pokud znalost o ontologii spočívá v tom, že máme k dispozici dva soubory znalostí $B = (\beta_1,\ldots,\beta_m)$ a $\Gamma = (\gamma_1,\ldots,\gamma_m)$ a odhad pravděpodobnostní distribuce stavů ontologie je q, potom pokud soubor znalostí B je z hlediska výše zavedené míry „informativnější“ než soubor znalostí Γ, potom znalosti B dovolují jednoduchou, výše zavedenou ϕ-transformací, transformovat distribuci q na distribuci, která je blíže k neznámé skutečné distribuci stavů ontologie p, než by to umožnily znalosti Γ. (Pokud žádný odhad skutečné distribuce stavů ontologie není k dispozici, předpokládáme, že výchozí distribuce q má rovnoměrné rozložení.)

Znalosti s větším informačním obsahem tedy dovolují přesnější aproximaci skutečného pravděpodobnostního rozložení stavů ontologie. Optimální rozhodování v rámci dané ontologie vyžaduje znalost skutečné distribuce stavů ontologie. Teorie s větším informačním obsahem tedy poskytují potenciální možnost lépe rozhodovat. Kromě toho je zřejmé, že zavedená míra informačního obsahu je v souladu s porovnáváním informativnosti na základě relace logického vyplývání (viz bod 2. výše uvedené věty) .

Příklad Předpokládejme, že nás zajímá souvislost mezi třemi příznaky pacienta a jeho možnou nemocí. Ontologii tedy tvoří pacient a jeho příznaky a nemoc, kterou může trpět. Jazyk J popisující ontologii bude obsahovat individuální konstantu $\boldsymbol{p}$ označující pacienta a predikáty π_1, π_2, π_3 označují příznaky pacienta a π_4 jeho nemoc. Dále budeme pro zjednodušení zápisu psát $\pi_1(\boldsymbol{p}) = a$, $\pi_2(\boldsymbol{p}) = b$, $\pi_3(\boldsymbol{p}) = c$, $\pi_4(\boldsymbol{p}) = d$.

Protože je $s=1$ a $r=4$, ontologie může být v jednom z 16 stavů. Konjunkce, které popisují jednotlivé stavy ontologie, lze snadno uspořádat:

$\alpha_0 = \neg a \wedge \neg b \wedge \neg c \wedge \neg d$

$\alpha_1 = \neg a \wedge \neg b \wedge \neg c \wedge d$

……

$\alpha_{15} = a \wedge b \wedge c \wedge d$.

Potom například formule $\beta = ((a \wedge b) \supset d) \wedge (d \supset (a \wedge b))$ je ekvivalentní formuli

$$\vee_{i\in I_\beta} \alpha_i, \quad \boldsymbol{I}_\beta = \{0, 2, 4, 6, 8, 10, 13, 15\}.$$

Předpokládejme, že je k dispozici velký soubor vyšetřených pacientů s ověřenou diagnózou a že chceme zjistit zda mezi symptomy a nemocí pacienta existují zákonitosti, které by bylo možné vyjádřit logickými formulemi. Použijeme proto data mining metodu generování asociačních pravidel a generujeme IF THEN pravidla.

Předpokládejme dále, že fyziologická podstata vztahů mezi symptomy a nemocemi je taková, že se mohou realizovat pouze následující tři kombinace příznaků a nemoci : $a\ b\ c\ d$, $\neg a\ \neg b\ \neg c\ \neg d$, $a\ \neg b\ c\ d$. Za této situace data mining může vést například ke stanovení následujících IF THEN pravidel:

$a \wedge b \supset d$ (tj. IF a AND b THEN d), $a \wedge c \supset d$, $\neg a \wedge \neg c \supset \neg d$, $a \supset d$, $b \supset d$, $\neg c \supset \neg d$

Na základě získaných IF THEN pravidel můžeme formulovat například tři následující teorie o vztahu symptomů a nemocí.

$T_1 = \{(a \wedge b) \supset d, (a \wedge c) \supset d, (\neg a \wedge \neg c) \supset \neg d\}$, $T_2 = \{a \supset d, \neg c \supset \neg d\}$, $T_3 = \{ a \supset d, \neg c \supset \neg d, b \supset d\}$

Vzniká otázka, která z těchto teorií je nejvíce informativní. K stanovení informačního obsahu těchto teorií použijeme výše zavedenou informační míru $I(T)$. Vzhledem k tomu, že nemáme k dispozici žádný odhad pravděpodobnostní distribuce jednotlivých stavů ontologie, budeme předpokládat rovnoměrné rozložení. Pokud spočteme informaci $I(T)$ pro jednotlivé teorie, dostaneme $I(T_1) = 0.37$, $I(T_2) = 0.69$, $I(T_3) = 0.98$. Je vidět, že teorie T_1 má menší informační obsah než teorie T_2, přestože obsahuje větší počet formulí. Informační obsah teorie T_3 je větší než informační obsah teorie T_2, což je v souladu s tím, že $T_2 \subset T_3$.

Při výpočtu informace $I(\beta_1, \ldots, \beta_m)$ se sčítají ty pravděpodobnosti q_i , které příslušejí stavům ontologie, ve kterých jsou znalosti $\beta_1, \ldots, \beta_m$ splněny (viz (2)). Je proto třeba projít všechny stavy, což samozřejmě není prakticky možné, pokud je počet možných stavů ontologie příliš veliký. Tato skutečnost tedy omezuje použití navržené informační míry v praktických aplikacích.

V oblasti dobývání dat je ale rámec vymezující ontologii často zvládnutelný. Jako příklad uveďme hledání produkčních pravidel, která dávají do souvislosti binární symptomy pacientů a jejich možné choroby. Ontologií jsou tady možné stavy pacientů charakterizované jejich symptomy $\pi_1, \ldots, \pi_r$, a jejich chorobami $\pi_{r+1}, \ldots, \pi_s$. Počet možných stavů ontologie je v tomto případě 2^s a výpočet informace podle (2) je výpočetně zvládnutelný pro řádově desítky symptomů a chorob.

LITERATURA

Berka, P., (2003): *Dobývání znalostí z databází. Praha*, Academia, 2003.

Cover, T. M., Thomas, J. A., (2006): *Elements of Information Theory*, John Wiley & Sons, New York.

Mendelsson, E., (1997): *Introduction to Mathematical Logic,* Chapman & Hall, London.

Vajda, I., Veselý, A., Zvárová, J., (2005): On the amount of information resulting from empirical and theoretical knowledge, *Rev. Math. Complutense*, **18**, 275-283.

Tato práce vznikla za podpory výzkumného záměru MSM 6046070904 „Informační a znalostní podpora strategického řízení“.