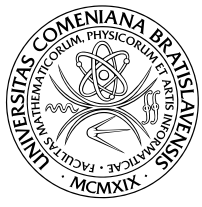
Fakulta matematiky, fyziky a informatiky
Univerzity Komenského v Bratislave



RNDr. Ľuboš Steskal

Autoreferát dizertačnej práce

On Usefulness of Information: a Computational Approach

na získanie vedecko-akademickej hodnosti philosophiæ doctor
v odbore doktorandského štúdia: 9.2.1. informatika

Bratislava 2010

Dizertačná práca bola vypracovaná v internej forme doktorandského štúdia na Katedre informatiky Fakulty matematiky, fyziky a informatiky Univerzity Komenského v Bratislave.

Predkladateľ: RNDr. Ľuboš Steskal
Katedra informatiky
Fakulta matematiky, fyziky a informatiky
Univerzity Komenského
Mlynská dolina
842 48 Bratislava

Školiteľ: prof. RNDr. Branislav Rovan, PhD.
Katedra informatiky FMFI UK
Bratislava

Oponenti: .....

.....

.....

Obhajoba dizertačnej práce sa koná dňa ..... o ..... h.
pred komisiou pre obhajobu dizertačnej práce v odbore doktorandského štúdia
vymenovanou predsedom odborovej komisie dňa .....

v študijnom odbore 9.2.1. informatika

na Fakulte matematiky, fyziky a informatiky UK, Mlynská dolina, miestnosť .....

Predseda odborovej komisie:
Prof. RNDr. Branislav Rovan, PhD.
Fakulta matematiky, fyziky a informatiky
Univerzity Komenského
Mlynská dolina
842 48 Bratislava

1 Úvod

Vedecký pokrok, najmä v oblasti telekomunikácií, viedol v minulom storočí k potrebe formálneho a matematického skúmania niektorých vlastností informácií. Prelomovou prácou v tejto oblasti je dielo Clauda Shannona [Sha48], v ktorom zaviedol pojmy ako kapacita kanálu, vzájomná informácia či entropia náhodného zdroja chápaná ako priemerná veľkosť informácie, alebo miera náhodnosti. Rozvíjaním jeho práce vznikla nová oblasť na pomedzí matematiky a informatiky – teória informácie.

Solmonoff [Sol64], Kolmogorov [Kol65] a Chaitin [Cha87] ďalej skúmali pojmy náhodnosti a veľkosti informácie z pohľadu teórie vypočítateľnosti, čo viedlo k vzniku Kolmogorovskej zložitosti, resp. algoritmickej teórii informácie.

Rozvoj a aplikácia poznatkov získaných z týchto teórií umožnili vytvorenie účinných komunikačných a telekomunikačných zariadení. Rovnako prispeli k lepšiemu pochopeniu pojmu náhodnosti a ponúkli pevný základ pre štúdium základných vlastností pojmu informácia.

V posledných rokoch sme však svedkami ďalšieho rozvoja, kedy automatizovaní agenti neplnia len úlohu sprostredkovateľa komunikácie, ale čoraz viac plnia aj rozhodovaciu úlohu. Možno o nich povedať, že efektívne manipulujú s dodanými informáciami a výsledkom ich práce je často opäť informácia. Skúmať vlastnosti takýchto agentov len z pohľadu komunikácie a ignorovať ich výpočtové obmedzenia ako aj obsah komunikovaných informácií zrejme ponúkne iba obmedzený vhľad do celej problematiky.

Táto situácia si žiada ďalší rozvoj v skúmaní vlastností informácií. Je zrejmé, že práve informatika (computer science) sa zaoberá mnohými otázkami a vlastnosťami informácie, ktoré štandardná teória informácie nepokrýva. Tieto skúmania však často pracujú s pojmom informácie len implicitne bez jednotnej formalizácie pojmu informácie.

V poslednom desaťročí však tieto a iné motivácie viedli viacerých vedcov k začatiu skúmania nových a doteraz explicitne neformulovaných problémov spojených s pojmom informácie. Sú tu snahy o zovšeobecnenú teóriu informácie [Top08] [Kåh02], modelovanie významu a užitočnosti informácie [GR08] a všeobecné porozumenie pojmu informácie [ATWG08].

V tejto práci ponúkame náš pohľad na pojem užitočnosti informácie. V prvej časti sa zaoberáme konceptuálnou analýzou tohto pojmu. Uvádzame, ako je informácia a najmä užitočnosť informácie videná a modelovaná v rôznych matematických a informatických teóriách. Snažíme sa poukázať na zistenie, že užitočná informácia nemusí byť nutne spätá so znížením neurčitosti. Ukážeme príklady, kedy užitočná informácia znižuje zložitosť riešenia problému. Ukážeme, že užitočnosť informácie sa dá vždy vnímať z pohľadu znižovania primeranej zložitostnej miery. Tieto pozorovania následne sformulujeme v polo-formálnom modeli informácie.

V ďalšej časti ponúkneme ukážku nášho prístupu a nášho modelu. Sformulujeme základné pojmy teórie užitočnosti informácie pre regulárne jazyky. Porovnáme dva intuitívne, avšak rôzne pohľady na nezávislosť regulárnych jazykov a ukážeme že rozdiel je spôsobený rôznou interpretáciou reprezentácie informácie agentom.

V poslednej časti sa zameriame na praktický problém automatickej detekcie plagiátov. Ukážeme, ako pomocou vypočítateľnej analógie normalizovanej informačnej vzdialenosti (ktorá je mierou užitočnosti informácie) možno ponúknuť uspokojivé riešenie skúmaného problému.

2 Ciele dizertačnej práce

Cieľom práce je skúmať užitočnosť informácie, najmä z algoritmického pohľadu, navrhnúť model užitočnosti informácie a zistené výsledky prípadne aplikovať v praxi, analyzovať rozličné prístupy k študovaniu informácie a jej užitočnosti a zistiť, či existujú situácie, kedy sa užitočnosť informácie nedá vnímať z pohľadu zmeny dĺžky optimálneho kódu. Ďalej je cieľom zadefinovať užitočnosť informácie pre triedu deterministických konečných automatov a skúmať vzťah medzi užitočnosťou a zložitosťou informácie.

3 Hlavné výsledky

3.1 Konceptuálny model

V práci skúmame viaceré matematické a informatické teórie, ktoré priamo či nepriamo narábajú s pojmom informácia. Ukazujeme, že v každom modeli možno identifikovať agenta, ktorý má nejaký cieľ, a užitočná informácia mu umožňuje efektívnejšie riešenie tohto cieľa. V každom modeli identifikujeme rozdiel medzi reprezentáciou informácie a jej sémantikou, rovnako ako aj rozdiel medzi problémom, ktorý agent rieši, a reprezentáciou problému.

Abstrakciou z uvedených prístupov dospievame k záveru, že každý z nich možno formulovať ako usporiadanú päticu $(\mathcal{I}, \mathcal{S}, \mathcal{P}, C, A)$, kde $\mathcal{I}$ je množina všetkých reprezentácií informácií, $\mathcal{S}$ je množina všetkých prípustných riešení, $\mathcal{P}$ je množina reprezentácií všetkých problémov, C je zložitostná funkcia a A je agent, ktorý na základe reprezentácie problému a informácie ponúkne riešenie.

Užitočnosť informácie I vzhľadom na problém P a agenta A možno potom definovať ako $U_A(I, P) = C_A(P) - C_A(P|I)$, teda rozdiel medzi zložitosťou riešenia (ponúknutého agentom) bez príslušnej informácie a s ňou.

3.2 Regulárne jazyky

Skúmame základné informačné vzťahy medzi regulárnymi jazykmi reprezentovanými úplnými minimálnymi deterministickými konečnými automatmi (DFA). Úlohu formulujeme vo formálnom rámci vytvorenom v predchádzajúcej kapitole. Skúmame vplyv dodatočnej informácie na veľkosť minimálneho automatu akceptujúceho jazyk P. Dodatočná informácia je reprezentovaná jazykom I s významom, že pre všetky slová w, ktoré sa môžu objaviť na vstupe, platí $w \in I$. Za predpokladu tejto informácie môže dôjsť k zjednodušeniu automatu potrebného na rozhodnutie, či $w \in P$.

V našej novej terminológii môžeme skúmanú úlohu reprezentovať usporiadanou päticou $(\mathcal{I}, \mathcal{S}, \mathcal{P}, C, A)$, kde $\mathcal{I} = \mathcal{P} = \mathcal{R}$, $\mathcal{S}$ je množina úplných deterministických automatov, $C(s, P|I) = C(s)$ je stavová zložitosť automatu s a agent A je

$$A(P|I) = \arg\min_{s \in \mathcal{S}} \{C(s); s \text{ rozpoznáva } P|I\},$$

pričom $P|I$ je promise problem $(P \cap I, P^C \cap I)$. Pod zložitosťou $C(P|I)$ rozumieme $C(A(P|I))$.

Nezávislosť dvoch regulárnych jazykov A a B definujeme dvoma spôsobmi:

1. A je podmienene nezávislé od B, ak $C(A) = C(A|B)$,
2. A a B sú spoločne nezávislé, ak $C(A \cap B) = C(A) \cdot C(B)$.

Ukazujeme, že ak sú jazyky A a B nezávislé v druhom zmysle, tak sú aj v prvom, ale nie nutne naopak.

V práci ďalej skúmame vzťah medzi $C(A|B)$ a $C(A \cap B)$ a ukazujeme, že platia podobné vlastnosti ako pre podmienenú pravdepodobnosť:

$$\frac{C(L_1 \cap L_2)}{C(L_2)} \leq C(L_1|L_2) \leq C(L_1 \cap L_2).$$

V práci tiež skúmame otázku symetrie informácie a ukazujeme, že existujú jazyky s výrazne asymetrickou užitočnosťou informácie.

Následne sa zaoberáme vzťahom medzi užitočnosťou informácie a jej zložitosťou. Ukazujeme, že ak má problém P zložitosť k, tak potom pre každé $l > O(k \cdot \log k)$ existuje jazyk I so zložitosťou l, ktorý má nulovú užitočnosť vzhľadom na problém P. Tým sme dokázali očakávané tvrdenie, že veľká zložitosť informácie ešte negarantuje užitočnosť.

Prekvapivo sme však našli jazyk $P' = \{a^{4k+3} | k \in \mathbb{N}\} \cup \{\varepsilon\}$, pre ktorý je každý jazyk so zložitosťou 2 užitočný.

3.3 Detekcia plagiátov

Ukazujeme nový prístup k problému detekcie dokumentových plagiátov, ktorý vychádza z teoretického pojmu normalizovanej informačnej vzdialenosti. Tento prístup sme implementovali a úspešne použili na reálnych dátach.

4 Zoznam použitej literatúry

[ABC+06] Rudolf Ahlswede, Lars Bäumer, Ning Cai, Harout K. Aydinian, Vladimir Blinovsky, Christian Deppe, and Haik Mashurian (eds.), *General theory of information transfer and combinatorics*, Lecture Notes in Computer Science, vol. 4123, Springer, 2006.

[ATWG08] P. Adriaans, P. Thagard, J. Woods, and D.M. Gabbay (eds.), *Philosophy of information*, North-Holland, 2008.

[BCB+09] C. Basile, G. Cristadoro, D. Benedetto, E. Caglioti, and M. Degli Esposti, *A plagiarism detection procedure in three steps: selection, matches and" squares"*, 3rd PAN WORKSHOP. UNCOVERING PLAGIARISM, AUTHORSHIP AND SOCIAL SOFTWARE MISUSE, 2009, p. 19.

[BGL+97] Charles H. Bennett, Péter Gács, Ming Li, Paul M.B. Vitányi, and Wojciech H. Zurek, *Information distance*, 1997.

[Cal02] C. Calude, *Information and randomness: An algorithmic perspective*, Springer Verlag, 2002.

[Cas07] E. Casanovas, *Simplicity simplified*, Revista Colombiana de Matemáticas **41** (2007), no. 1, 263–277.

[CFL+04] X. Chen, B. Francia, M. Li, B. McKinnon, and A. Seker, *Shared information and program plagiarism detection*, IEEE Transactions on Information Theory **50** (2004), no. 7, 1545–1551.

[Cha87] Gregory J. Chaitin, *On the length of programs for computing finite binary sequences: Statistical considerations, journal of the ACM, 1969*, Information Randomness & Incompleteness: Papers on Algorithmic Information Theory, Gregory J. Chaitin, World Scientific, Series in Computer Science– Vol. 8, 1987.

[CT06] Thomas M. Cover and Joy A. Thomas, *Elements of information theory, second edition*, John Wiley and Sons, Inc., 2006.

[DBL79] *Conference record of the eleventh annual acm symposium on theory of computing, 30 april-2 may, 1979, atlanta, georgia, usa*, ACM, 1979.

[ESY84] S. Even, AL Selman, and Y. Yacobi, *The complexity of promise problems with applications to public-key cryptography*, Information and Control **61** (1984), no. 2, 159–173.

[GKB+08] Viliam Geffert, Juhani Karhumäki, Alberto Bertoni, Bart Preneel, Pavol Návrat, and Mária Bieliková (eds.), *Sofsem 2008: Theory and practice of computer science, 34th conference on current trends in theory and practice of computer science, nový smokovec, slovakia, january 19-25, 2008, proceedings*, Lecture Notes in Computer Science, vol. 4910, Springer, 2008.

[GR08] Peter Gazi and Branislav Rovan, *Assisted problem solving and decompositions of finite automata*, in Geffert et al. [GKB+08], pp. 292–303.

[HR92] Jr. Hartley Rogers, *Theory of recursive functions and effective computability*, The MIT Press, 1992.

[HT06] Peter Harremoës and Flemming Topsøe, *Zipf's law, hyperbolic distributions and entropy loss*, in Ahlswede et al. [ABC+06], pp. 788–792.

[HU79] J. E. Hopcroft and J. D. Ullman, *Introduction to automata theory, languages and computation*, Addison-Wesley, 1979.

[Huf52] D. A. Huffman, *A method for the construction of minimum-redundancy codes*, Proceedings of the IRE **40** (1952), no. 9, 1098–1101.

[JB03] E.T. Jaynes and G.L. Bretthorst, *Probability theory: the logic of science*, Cambridge University Press Cambridge, 2003.

[Kåh02] Jan Kåhre, *The mathematical theory of information*, Kulwer Academic Publishers, London, 2002.

[KBK09] J. Kasprzak, M. Brandejs, and M. Kripac, *Finding Plagiarism by Evaluating Document Similarities*, 3rd PAN WORKSHOP. UNCOVERING PLAGIARISM, AUTHORSHIP AND SOCIAL SOFTWARE MISUSE, 2009, p. 24.

[Kol65] Andrey N. Kolmogorov, *Three approaches to the quantitative definition of information.*, Problems in Information Transmission **1** (1965), 1–7.

[LCL+04] M. Li, X. Chen, X. Li, B. Ma, and P.M.B. Vitányi, *The similarity metric*, IEEE Transactions on Information Theory **50** (2004), 12.

[LV08] Ming Li and Paul M.B. Vitnyi, *An introduction to kolmogorov complexity and its applications*, Springer Publishing Company, Incorporated, 2008.

[MKZ06] H. Maurer, F. Kappe, and B. Zaka, *Plagiarism-a survey*, Journal of Universal Computer Science **12** (2006), no. 8, 1050–1084.

[ML09] J.A. Malcolm and P.C.R. Lane, *Tackling the PAN'09 External Plagiarism Detection Corpus with a Desktop Plagiarism Detector*, 3rd PAN WORKSHOP. UNCOVERING PLAGIARISM, AUTHORSHIP AND SOCIAL SOFTWARE MISUSE, 2009, p. 29.

[MZK+09] M. Muhr, M. Zechner, R. Kern, M. Granitzer, and K.C. Graz, *External and Intrinsic Plagiarism Detection Using Vector Space Models*, Stein et al.(Stein et al., 2009) (2009).

[Ris07] Jorma Rissanen, *Information and complexity in statistical modeling (information science and statistics)*, Springer, January 2007.

[Sel09] J. Seligman, *Channels: From Logic to Probability*, Formal Theories of Information (2009), 193–233.

[Sha48] C. E. Shannon, *A mathematical theory of communication*, Bell Sys. Tech. J. **27** (1948), 379–423, 623–656.

[Sol64] R. J. Solomonoff, *A formal theory of inductive inference. part 1 and part 2.*, Information and Control **7.** (1964).

[Top08] Flemming Topsøe, *Game theoretical optimization techniques inspired by information theory.*

[Tsa88] C. Tsallis, *Possible generalization of Boltzmann-Gibbs statistics*, J. Stat. Physics **52** (1988).

[VBCL08] P.M.B. Vitanyi, F.J. Balbach, R.L. Cilibrasi, and M. Li, *Normalized information distance*, Information Theory and Statistical Learning (2008), 45–82.

[Yao79] Andrew Chi-Chih Yao, *Some complexity questions related to distributive computing (preliminary report)*, in *STOC* [DBL79], pp. 209–213.

[YZ91] S. Yu and Q. Zhuang, *On the state complexity of intersection of regular languages*, ACM SIGACT News **22** (1991), no. 3, 52–54.

[ZHZ+07] X. Zhang, Y. Hao, X. Zhu, M. Li, and D.R. Cheriton, *Information distance from a question to an answer*, Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2007, p. 883.

[ZL77] J. Ziv and A. Lempel, *A universal algorithm for sequential data compression*, IEEE transactions on Information Theory **23** (1977), no. 3, 337–343.

5 Summary

In this thesis we deal with the problem of useful information from an algorithmic perspective. Various approaches to the study of information and information usefulness are presented. We argument that the concept of information usefulness is not only about the reduction of the best code length. We show that it should be rather interpreted as the reduction of a complexity measure of a problem being solved.

We offer an example of this approach by studying basic information usefulness properties of regular languages. After showing that there are two intuitive, but different notions of language independence, we examine the relation between complexity and usefulness. We show that however complex the information, it does not necessarily help, but on the other hand that there are cases, where any simple enough information can help.

Finally, we present a real life application of the theoretical notion of normalized information distance. We use a computable analogue to construct an automated plagiarism detection tool.

Keywords: useful information, complexity, information theory, Kolmogorov complexity, normalized information distance, deterministic finite automata, promise problems, plagiarism detection.