



Informace, kódování, data

Dušan Saiko, FD ČVUT, K620

pro předmět Telematika

16.03.2010

saiko@lss.fd.cvut.cz

Představení

- Subjekt
 - základ práce každého informatika
 - zajímavé technické i filozofické poznatky
 - obecné využití znalostí
 - aplikace v dopravě
 - cíle přednášky
- Osoby
 - vzdělání:
 - 1997-2001
 - různá studia na ČVUT
 - 2006 UHK FIM, Informační management
 - praxe:
 - od roku 1997
 - od roku 2004 návrh a vývoj J2EE aplikací

Obsah

- Informace
- Data
- (Pre)Historie počítačů
- Kódování
- Entropie
- Komprese dat
- Šifrování

Informace

- Co to je informace
 - informace, lat. zpráva, popř. její obsah
 - Sdělení poznatku
 - veškerý technologický i sociální vývoj je postaven na shromažďování, předávání a vyhledávání informací
 - Claud Elwood Shannon
 - "Matematická teorie komunikace" [Shannon 1948], Bell System Technical Journal 1948
 - vymezuje jasně počátek teorie informace
 - ne však počátek používání a kódování informací

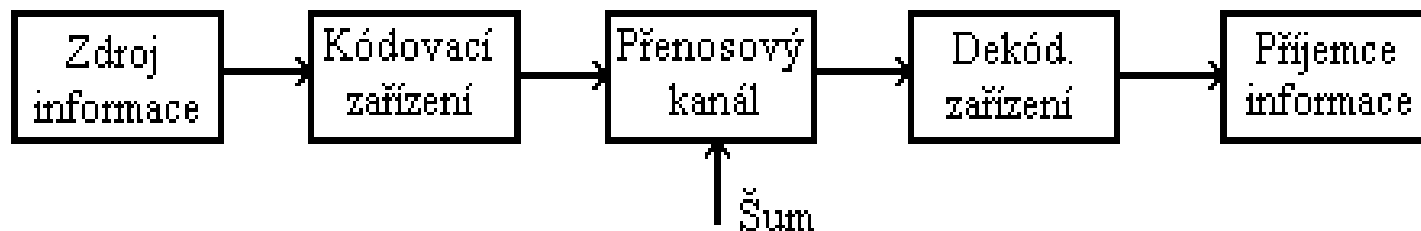
Informace

- Vlastnosti informace
 - počátek
 - záblesk světla, slovo, mechanický jev
 - šíření
 - nosičem (signálem)
 - při přenosu a uchovávání může být rušena
 - snižování kvality (opisy, "drby"), šum
 - má svého adresáta, který informaci využívá
 - konkrétní či obecný adresát

Informace

Informace

- vše co vnímáme (světlo, zvuk, ...)
- elektromagnetické záření



(Pre) Historie počítačů

- cca 3000 př.n.l.
 - Abakus - nástroj pro usnadňování počítání čísel
- 1623 - první mechanický kalkulátor
- 1833 - Charles Babbage - "Analytical Engine"
- 20 století - první mechanické stroje na šifrování zpráv
 - a naopak stroje (návrh počítačů) na dešifrování zpráv
 - Anglie během 2 sv. války, (viz též Alan Turing)

Data

- Data
 - zavedené pojmenování
 - data - obálka, v které jsou uloženy informace
 - informace - obsahuje určitou znalost

Data

- Číselné soustavy
 - desítková soustava 0....10
 - binární soustava 0..1
 - libovolná soustava
 - uživatelsky definovaná

Data

- Číselné soustavy
 - přičítání v libovolné soustavě:
 - poslední znak/číslo se zvýší o jedna na vyšší znak, pokud je poslední znak nejvyšší hodnota, sníží se na počáteční znak a stejným postupem se přičítá znak předchozí (významově vyšší)
 - dochází k vygenerování všech možných čísel - tedy všech možných kombinací dané číselné soustavy (znakové sady)

Data

- Převody mezi číselnými soustavami
 - např. binárně $10010 = \text{decimálně } 2^4 + 2^1$
 - omezení lidské představivosti pro práci s desítkovou soustavou
 - zajímavé vyzkoušet si samostatně doma vytvořit funkci pro převod mezi libovolnými číselnými soustavami

Kódování

- Obecný význam kódování:
 - převod informace do dat
- Příklady kódování
 - jazyk (řeč)
 - písmo (slova, věty)
 - uchování informace (fotografie)

Entropie

- Definice
 - míra neuspořádanosti
 - číselné vyjádření množství informace, které je průměrně obsaženo v datech
- Lze počítat
- Čím je v datech více informace, tím jsou data efektivněji využívána

Generování informace

- Filozofická otázka:
 - všechny informace již existují, jediným problémem je najít je v obrovském objemu šumu
 - (má vlastní úvaha)
- Funkce
 - pokud máme funkci, algoritmicky generující veškeré možné stavy určité omezené informace, pak každá informace může být vyjádřena jako index (číslo - pořadí výsledku) dané funkce

Generování informace

- Příklad - text:
 - anglická abeceda: 28 znaků + formátovací znaky (.,?!-), dle ASCII tabulky maximálně 128 viditelných znaků
 - generujeme všechny možné texty např. s omezením délky na 100 znaků
 - cca $5 \cdot 10^{210}$ možností, které ale obsahují všechny možné existující věty do 100 znaků zapsatelné v anglické abecedě.
 - 128^{100}

Generování informace

- Příklad - text:
 - generování textu naší funkcí
 - $f(1)$: a
 - $f(2)$: aa
 - $f(n1)$: baaaaaaaa
 - $f(n2)$: aaaefgadfafdsxc...
 - $f(n3)$: naprikład tento text
 - všechny možné existující texty do 100 znaků lze v návaznosti na daný algoritmus odkazovat indexem (pořadovým číslem) hodnoty funkce $f()$, kde $0 < \text{index} < 5 * 10^{210}$

Generování informace

- Příklad - text:
 - generování názvů (firmy)
 - generování všech možných kombinací písmen anglické abecedy, např. s rozsahem 5 až 7 znaků
 - aplikace pravidel na generované kombinace
 - střídání souhlásek a samohlásek
 - neopakují se za sebou dvě stejná písmena
 - omezení generovaných možností
 - překvapivé a využitelné výsledky
 - s tím, že většina "hezkých" názvů je již použita

Generování informace

- Příklad - text:
 - generování "přirozeného" textu
 - aplikování pravidel pro text v daném jazyce
 - statistika výskytu dvojic jednotlivých písmen
 - statistika délky slov
 - zajímavé výsledky, výsledek bez významu, ale s podobností přirozenému textu

Generování informace

- Příklad - počítačová ikona
 - $16 * 16$ bodů (256)
 - černobílá:
 - $2^{256} = 1,15.. * 10^{77}$
 - barevná (256 barev):
 - $256^{256} = 3,2... * 10^{216}$
 - návrh počítačem tvořené grafiky:
 - aplikace různých grafických pravidel

Morseova abeceda

- Morse, USA 1837
 - ternární kód (dlouhý zvuk, krátký zvuk, mezera)
 - navrhl abecedu tak, aby její zapsání bylo efektivní
 - nejpoužívanějším znakům navrhl nejkratší kódy
 - a .-
 - e .
 - využitím kódování informace předběhl dobu

Komprese dat

- Komprese
 - snaha o efektivní kódování informace
 - známý (popsaný) algoritmus, bez snahy utajit data
 - překlad již existujících dat do efektivnější podoby
 - dobře komprimovaná data - velká entropie - velká neuspořádanost dat - velký obsah informace na datovou jednotku - neuspořádanost dat - možné využití v náhodných číslech

Kompresa dat

- Kompresa
 - obecná
 - speciální
 - bezstrátová
 - ztrátová
 - dle druhu dat

Kompresa dat

- Další rozdělení
 - Účel:
 - přenos dat (záleží na rychlosti)
 - dlouhodobá archivace (může používat náročný algoritmus, cílem jsou co nejmenší data)
 - transparentní komprimace (např. diskových svazků - musí umožňovat co nejrychlejší přístup na jakoukoliv pozici v datech)

Kompresa dat

- Další rozdělení
 - Dle návrhu kompresních algoritmů
 - statistické algoritmy
 - pracují se známými statistikami dat či jeho zdroje
 - slovníkové algoritmy
 - udržují slovník vyskytujících se spojení
 - analytické algoritmy (pro ztrátovou kompresi)
 - nalézají matematické funkce pro prezentaci obsahu dat
 - aproximační metody (pro ztrátovou kompresi)
 - aproximace hodnot
 - numerické metody, neuronové sítě
 - v praxi prolínání metod, vícenásobná komprese (např bezztrátová komprese aplikovaná na výsledek ztrátové komprese)

Kompresa dat

- Základní algoritmy
 - RLE - Run Length Encoding
 - populární v historii formátů počítačové grafiky pro soubory PCX, BMP, TIFF
 - rychlá, jednoduchá
 - nevalné výsledky pro složitější a více neuspořádanější obsah
 - nevhodné pro obecná data

Kompresa dat

- Základní algoritmy
 - RLE - Run Length Encoding
 - příklad: obrázek s jednobarevným pozadím (dokument)
 - jednorozměrný vstup může být zapsán jako výstup množin o dvou prvcích, první značí počet opakování, druhý konkrétní prvek (barvu)
 - vstup: AAAAAABBBBCCCC (14 znaků)
 - výstup: 6A4B3C14 (8 znaků)
 - kompresní poměr (zkomprimováno na): 57 %
 - dosažená komprese: 43 %

Kompresa dat

- Základní algoritmy
 - Diatomické kódování
 - např pro jednoduchou kompresi textu
 - vyhledání znaků, které se v textu nevyskutují a jejich použití pro nahrazení nejčastěji vyskytujících se dvojic znaků v textu
 - vícenásobné použití

Kompresa dat

- Základní algoritmy
 - Diatomické kódování
 - vstup: AAAAAAAAAA
 - první nepoužitý znak: 1, nejčastější dvojice: AA
 - výstup1: 11111
 - další nepoužitý znak: 2, nejčastější dvojice: 11
 - výstup2: 221
 - další nepoužitý znak: 3, nejčastější dvojice: 22
 - výstup3: 31

Kompresa dat

- Huffmanovo statistické kódování
 - známe statistiky vstupních dat
 - jednotlivým znakům (skupinám) přiřadíme kód dle výskytu pravděpodobnosti, např.
 - A 1
 - B 01
 - C 001
 - v historii komprese důležitý algoritmus

Kompresa dat

- LZW/LZ77/LZ78
 - další důležitý algoritmus
 - Lempel-Ziv-Welch
 - ZIP ARJ GZIP
 - hledá shodné skupiny znaků v datech, nahrazuje je odkazem na předchozí výskyt skupiny

Kompresa dat

- Transformace dat před kompresí
 - úprava dat tak, aby šla dobře komprimovat
 - Delta encoding
 - pro data s charakteristikou funkce
 - nahrazení dat inkrementy mezi jednotlivými znaky (číslý)
 - vstup: 0123456789
 - výstup: 0111111111

Kompresa dat

- Transformace dat před kompresí
 - segmentace hladin hodnot
 - nahrazuje segmenty dat s podobnou hodnotou daty s odečteným základem společné hodnoty
 - vstup: 255 240 230 254
 - výstup: 25 10 0 24 (vstup - 230)
 - vstup: 255 240 230 254 20 50 10
 - výstup 25 10 0 24 | 10 40 0
 - zajímavá transformace, složitá na vyhledání optimálního seskupení dat

Kompresa dat

- Burrows-Wheelerova transformace
 - bzip2
 - významný algoritmus
 - nahrazení dat jednodušší reprezentací s minimální přidanou hodnotou
 - vytvoření všech rotací vstupního řetězce
 - setřídění rotací dle abecedy
 - výstup: poslední znaky setříděného listu

Kompresa dat

- Moderní kompresní algoritmy:
 - praxe:
 - LZMA
 - teorie:
 - neuronové sítě
 - quantová informatika

Šifrování

- Kódování dat s důrazem na nemožnost přečtení dat třetí osobou
- Velký historický význam
- Kryptoanalýza - nauka o dešifrování a prolamování šifer
 - používaná například i pro luštění neznámých písem

Šifrování

- Historie šifrování
 - první kryptografie již před letopočtem
 - kriptování tajných válečných zpráv
 - kriptování královské konverzace
 - poprava Marie Stuartovny 1586
 - ve 20tém století roste důležitost informací a jejich utajení
 - 2 sv válka - Enigma
 - podcenění nebezpečí války a šifrování po 1 sv válce

Šifrování

- Historie šifrování
 - dnes nedílná součást informatiky
 - asymetrické šifrování
 - objev veřejného a privátního klíče
 - matematické funkce založeny na dvou velkých prvočíslech
 - hrozba quantové kryptoanalýzy
 - velice zajímavý historický vývoj, fascinující příběhy lidí, kteří se šifrování/kryptoanalýze věnovali

Šifrování

- Historie šifrování
 - Utajení informací: velice důležité/nebezpečné
 - Philip Zimmermann
 - po celá 90tá léta vyšetřován kvůli uveřejnění programu PGP
 - Skutečný stav kryptoanalýzy známe s cca. 30 letým zpožděním

Šifrování

- Základní rozdělení šifrování
 - šifrování s utajeným algoritmem
 - dnes již jen jako dětské hry ...
 - šifrování s veřejným algoritmem a utajeným klíčem
 - bezpečné, vzniká problém přenosu klíče
 - šifrování s veřejným algoritmem a privátním/veřejným klíčem

Šifrování

- Základní rozdělení šifrování
 - šifrování s utajeným algoritmem
 - dnes již jen jako dětské hry ...
 - šifrování s veřejným algoritmem a utajeným klíčem
 - bezpečné, vzniká problém přenosu klíče
 - šifrování s veřejným algoritmem a privátním/veřejným klíčem
 - první vynalezení 1973, první veřejná publikace 1978, veřejné použití - druhá polovina 90tých let

Šifrování

- Šifrování s utajeným klíčem

- problém distribuce klíčů (např. ve válce pro Enigmu), hledání řešení problému
- úvaha:
 - Odesílatel pošle skříňku opatřenou zámkem A, klíč si uschová. Příjemce nemůže otevřít skříňku, nemá klíč. Příjemce zamkne skříňku druhým zámkem B a pošle zásilku zpět. Původní odesílatel odemkne zámek A a znovu pošle příjemci. Příjemce odemkne zámek B a může otevřít skříňku.
 - dvojité posílání zprávy
 - nutná aktivní komunikace příjemce/odesílatele (nelze něco zabezpečeného poslat na pasivní adresu)
 - problémy s hledáním vhodného šifrovacího algoritmu, který by umožňoval po sobě zašifrování 1, zašifrování 2, dekodování 1, dekodování 2
 - » dle mého úsudku lze použít nejjednodušší (ale s kvalitním klíčem neprolomitelné) šifrování XOR
- později vyřešeno algoritmem s veřejným a privátním klíčem

Šifrování

- Steganografie
 - ukrývání zpráv
 - historie: neviditelné inkousty, ukrývání zpráv pod vlasy
 - moderní: ukrývání dat v hudebních souborech ...
- Transpoziční šifrování
 - změna pořadí písmen
- Substituční šifrování
 - nahrazení písmen / skupin jiným kódem
- Viz literatura

Doporučená literatura

- Kniha kódů a šifer
 - Simon Singh
 - Tajná komunikace od starého Egypta po kvantovou kryptografii
 - odborná literatura s prvky beletrie
 - můj názor: základní literatura každého informatika
- Vyhledávač na klíčová slova
 - Information Theory and Coding
 - history of cryptography
 - history of computers
 - compression algorithms
- Existující diplomové a bakalářské práce pro daný okruh témat

Konec

- Děkuji za pozornost
- Dotazy, připomínky
- Semestrální práce

