Masarykova univerzita
Fakulta informatiky

Power law

MV011 Statistika I – Projekt

Jiří Kremser

Úvod

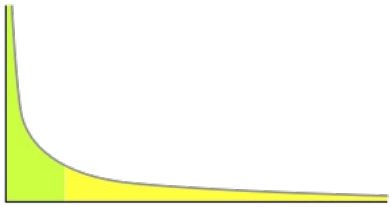
Ve svém projektu se budu zabývat tzv. power law schématem (překládá se jako mocninný zákon). Jedná se o častý vzor vyskytující se ať už v přírodě člověkem netknuté, či v oblastech ryze vykonstruovaných lidmi. Neformálně by se dalo tvrdit, že power law je typický pro objekty/jevy s nějakou vnitřní strukturou, zatímco normální zákony pro objekty/jevy s chaotickou strukturou [1] a také, že graf power law distribuce má tvar podobný hyperbole, zatímco normální zákony podléhají většinou rozložení tvaru nějaké Gaussovy křivky. Pro lepší pochopení problematiky uvedu ještě před definicí několik jevů, které podléhají power law, z nichž bude podstata power law mnohem jasnější.

Instance

- **Zipfův zákon**
  Je empirické pravidlo pocházející z oblasti lingvistiky, které využívá matematickou statistiku k počítání četnosti slov. Relativní četnost slov v textu splňuje Zipfovo rozložení [2], což je speciální případ obecnějšího power law rozložení. Situace je tedy taková, že pouze několik málo slov má relativní četnost enormně vysokou, zpravidla se jedná o členy, předložky, zájmena apod. Je zřejmé, že výsledek je závislý na referenčním korpusu a dostaví se až u dostatečně dlouhých textů. Mohlo by se zdát, že relativně častá slova jsou klíčová v daném textu. Určování klíčovosti (keyness) a „tématu“ (aboutness) textů je disciplínou zpracování přirozeného jazyka a je prováděno mnohem sofistikovanějšími metodami.

- **Pravidlo 80-20**
  Známé též jako Paretův princip je nejčastěji zmiňováno s ekonomikou či kvalitou výroby. Lze jej formulovat tak, že 80 % důsledků (např. zisk nebo počet zmetků) pramení z 20 % příčin (např. produkty nebo celková výroba). V praxi potom bývá snahou odhalit ono malé spektrum příčin, které tak významně ovlivňuje celkový výsledek. Tento proces hledání se nazývá Paretova analýza. Princip lze rovněž aplikovat na příjmy zemí světa, následující tabulka zachycuje stav z roku 1989 (je převzata z [3]).

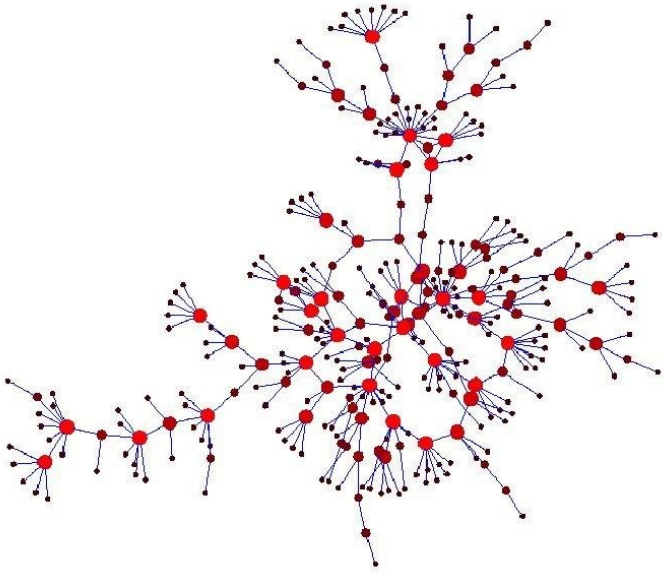
Část obyvatelstva	Příjem
Nejbohatších 20 %	82.7%
Druhých nejbohatších 20 %	11.7%
Třetích nejbohatších 20 %	2.3%
Čtvrtých nejbohatších 20 %	1.4%
Nejchudších 20 %	1.2%



Graf pro pravidlo 80 - 20

- **Citační síť**
  V modelu, kde si vizualizujeme publikace jako vrcholy grafu a hrana se vykytuje mezi dvěma vrcholy právě tehdy, citovala-li jedna publikace tu druhou, můžeme pozorovat opět power law. Stupně vrcholů (počet hran z vrcholu vycházející) v takto vytvořeném grafu totiž podléhají power law distribuci. Pro zjednodušení uvažujme graf s neorientovanými hranami. Je pochopitelné, že většina všech citací je směřována ke „slavným" dílům. Citovanost (počet citací) tak teoreticky může určovat přínos či prestiž publikace. Takovýchto nejcitovanějších[1] publikací je málo. Naopak publikace typu bakalářské práce většinou moc citované nejsou, ale je jich relativně mnoho. Je zajímavé, že v sítích, jejichž stupně vrcholů podléhají distribuci power law (nazývají se scale-free [4]) se častěji vyskytuje fenomén tzv. „malého světa". Jedná se o jev, kdy lze 2 libovolné vrcholy grafu/sítě spojit nejvýše konstantním počtem hran, nebo neformálně řečeno „hopů" přes vrcholy. Například v síti matematických publikací se jedná o tzv. Erdovo číslo (Erdős number) [5].

- **Síť sexuálních styků**
  Vrcholy představují lidi a hrana vede mezi dvěma vrcholy, měli-li dva dotyční pohlavní styk. Stupně vrcholů podléhají distribuci power law. Graf může usnadnit prevenci pohlavních chorob, je však prakticky nezískatelný.



Ukázkový graf sexuálních styků [5]

- **Jak dlouho žije druh**
  Tento příklad, jako jediný nesouvisí se systémem založeným člověkem. Drtivá většina druhů, které kdy žily na Zemi, vymřela. David Raup zformuloval [6] matematické modely, které mají jako společnou vlastnost výraznou asymetrii. Pro dobu života druhu neplatí klasická Gaussova křivka ve stylu "polovina všech druhů žije kolem milionu let a na obě strany od této hodnoty četnost klesá". Raupovy modely naopak tvrdí něco ve smyslu "téměř všechny druhy žijí velmi krátce a jen několik málo druhů žije velmi dlouho".

1 Tipuji, že nejcitovanější je bible.

Definice vztahu power law

Formálně se jedná o polynomiální funkci mající tvar $f(x) = ax^k + o(x^k),$
kde o(x^k) je funkce proměnné, která roste asymptoticky pomaleji než x^k .

Funkce *f* musí být soběpodobná[2] (*scale invariance*), to znamená, že platí vztah
$f(cx) = a(cx)^k = c^k f(x) \propto f(x)$. *c* je konstanta. Po vynásobení argumentu funkce *f* konstantou *c*, má tedy graf takovéto funkce stejný tvar jako funkce s původním argumentem, až na posun. Neformálně by se dalo říct, že nezáleží na měřítku.

Definice distribuce power law

Řekneme, že pravděpodobnostní rozložení je power law, jestliže jeho pravděpodobnostní funkce / hustota je tvaru $p(x) = \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\alpha}$
kde $x^{-\alpha}$ je normalizační konstanta a x_{min} je hranice, od které platí power law.

Speciálním případem power law rozložení je například nejjednodušší zeta rozložení s pravděpodobnostní funkcí (diskrétní).

$$\pi\{x = k; a\} = \frac{k^{-a}}{\zeta(a)}$$

kde $a > 1$ je parametr rozložení a $\zeta(a) = \sum_{i=1}^{\infty} \frac{1}{i^a}$ je Riemannova zeta funkce, která slouží jako normalizační konstanta [7].

Následující pravděpodobnostní rozložení jsou také power law:

- Paretovo rozložení
- Yule–Simonovo rozložení
- Zipfovo rozložení
- Studentovo t-rozložení

Parametr *a* power law rozložení byl empiricky zjištěn [8] pro následující sítě. Respektive pro rozložení stupňů vrcholů v těchto sítích následovně:

- Internet a WWW: $a = 2{,}1$
- Sociální sítě (např. citační grafy): $a = 3$
- Biologické sítě (grafy interakce proteinů): $a = 2{,}5$

Většina grafů s power law distribucí stupňů vrcholů reprezentujících reálný svět má $a \in [1, 4]$.

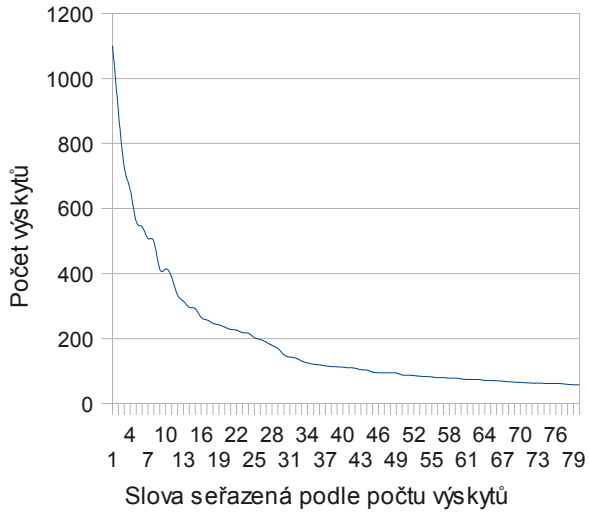
2 Soběpodobnost je charakteristická pro fraktály, kde libovolným zvětšením objektu dostáváme objekt původní.. Slevíme-li z rigidní definice, můžeme podobné rysy pozorovat také v přírodě např. struktura kapradí, mrak a jiné.

Praktická část

Provedl jsem analýzu dramatu Hamlet od Williama Shakespeara. Drama obsahuje v původní verzi 32241 slov. Po odstranění scénáristických informací zbylo 29968 slov [10]. Z tohoto souboru, jsem vybral 80 nejfrekventovanějších. Kvůli archaické angličtině se v první osmdesátce vyskytují dnes už zastaralé výrazy jako *„thou“*, *„thy“* a jiné. To ale není zajímavé. Zajímavější je fakt, že nejfrekventovanějších 80 slov v celém díle zabírá více než polovinu slov, přesně 15519 (15519/29968 = 51,8 %). Vynesl jsem získaná slova do tabulky a grafu níže. Vzhledem k faktu, že graf zachycuje pouze prvních 80 slov, není na něm vidět „dlouhý ocásek“ (long tail [10]), který je typický pro power law. Na získaném souboru spočítám střední hodnotu a rozptyl. Tyto údaje nebudou přesné, protože zanedbám zbylá slova (z „ocásku“), pro účely tohoto projektu by to ale snad mohlo stačit.

80 nejfrekventovanějších slov v Hamletovi

The	1101	That	389	As	228	No	143	All	110	Thy	87	Know	74	Very	64
And	898	Is	334	Be	226	We	140	Good	109	Her	86	Sir	74	Speak	63
To	726	Not	315	Lord	218	Are	131	Come	104	At	84	Them	74	Which	63
Of	657	This	296	He	216	On	125	Thou	103	Was	83	May	71	Hath	62
I	561	His	292	What	203	O	121	Now	97	Most	82	Tis	71	Then	62
You	544	But	265	So	197	Our	119	From	95	Like	80	Go	70	Why	62
My	508	With	257	Him	189	By	116	More	95	Would	80	Us	69	Must	61
A	498	For	247	Have	179	Shall	114	They	95	Hamlet	78	King	67	Thee	59
In	414	Your	242	Will	169	If	113	Let	94	Well	78	Love	66	Give	58
It	414	Me	235	Do	150	Or	112	How	88	There	76	Did	65	Should	58



Střední hodnota a rozptyl vybraného souboru slov

Jev X bude, že náhodně vybrané slovo z Hamleta, kde odstraníme všechna méně frekventovaná slova (81. nejfrekventovanější a dále.), bude X-té nejfrekventovanější.
Tedy:

$$\pi(1)=1101/15519\approx 0{,}0709$$
$$\pi(2)=898/15519\approx 0{,}0578$$
$$\pi(3)=726/15519\approx 0{,}0467$$
$$\pi(4)=657/15519\approx 0{,}0423$$
$$\vdots$$
$$\pi(79)=\pi(80)=58/15519\approx 0{,}0037$$
$$\pi(x)=0 \; pro \; x\in\mathbb{N}\setminus\{1..80\}$$
$$\boldsymbol{E(X)}=\sum_{x=1}^{80} x\pi(x)=0{,}0709+0{,}1157+0{,}1403+0{,}1693+...+0{,}2990\approx \mathbf{22{,}4}$$
$$\boldsymbol{D(X)}=E(X^2)-[E(X)]^2=\sum_{x=1}^{80} x^2\pi(x)-22{,}4^2=0{,}0709+0{,}4629+1{,}2631$$
$$+...+1913{,}5254\approx 51115{,}08-601{,}76=\mathbf{50513{,}32}$$

Kompletní výpočet je v přiloženém souboru *Hamlet-top80.ods*. Jde vidět, že rozptyl roste s každým dalším x-tým nejfrekventovanějším slovem, proto pro spočetně mnoho slov by neexistoval. Ve spojitém případě také ne.

Závěr

Ve svém projektu jsem se zaměřil na fenomén power law, který se obvykle vynoří v lidmi vytvořených systémech, to ale není pravidlem viz příklad s délkou života živočišných druhů. Nejzajímavější oblastí, kde se power law vyskytuje jsou sociální sítě. Možná by stálo za to provést analýzu sítí jako Facebook, či MySpace, které se řídí mocninným zákonem také. To ale nebylo možné pro nedostatek dat z těchto a jiných sítí. Využil jsem toho, že mám k dispozici Hamleta v originále a prezentoval power law na příkladu častých slov v tomto dramatu. V předmětu strojové učení pracujeme na projektu, v němž hledáme v Hamletovi časté vzory, ale sofistikovaněji. Jedním takovým častým vzorem je například *„my lord“*. V Hamletovi je *„my lord“* používáno poddanými k oslovení svého pána a také k oslovení Boha. Slova jako *„the“*, *„a“*, *„and“* a podobné, jsou cíleně odstraňována, protože zpravidla nic neříkají o podstatě daného díla. Doufám, že jsem nesklouzl až k příliš vágním spekulacím v úvodu a závěru a že projekt splní zadání i přes absenci náročnějších výpočtů.

Zdroje

[1] http://scienceworld.cz/technologie/mocninne-versus-normalni-zakony-a-co-z-toho-vyplyva-1716
[2] http://en.wikipedia.org/wiki/Zipf's_law
[3] http://en.wikipedia.org/wiki/80-20_rule
[4] http://en.wikipedia.org/wiki/Scale-free_network
[5] http://www.fi.muni.cz/~xpelanek/IV109/slidy/networks.pdf
[6] David Raup: O zániku druhů, Praha, 1995
[7] http://behind-the-enemy-lines.blogspot.com/2008/01/misunderstandings-of-power-law.html
[8] http://is.muni.cz/el/1433/podzim2008/PB165/um/nove10_2008.pdf?fakulta=1433;obdobi=4384;kod=PB165
[9] http://www.fi.muni.cz/~xkremser/hamlet.txt
[10] http://en.wikipedia.org/wiki/The_Long_Tail