Liet. matem. rink., T. 41, spec. nr., 430–437

Mokslinės terminijos modelių eksperimentinis tyrimas

Vaidas BALYS, MII
el. paštas: vbalys@delfi.lt

1. Įvadas

Labai sparčiai besivystant mokslo sritims iškyla natūralus poreikis išmokti įvertinti ir palyginti skirtingų šakų plėtimosi greitį, nustatyti dėsningumus ar net sudaryti prognozes ateičiai. Norint tai padaryti tenka stebėti mokslinius produktus – įvairias publikacijas, registruoti raktinius žodžius ir frazes, analizuoti jų aibės kitimo sąvybes, skaičiuoti įvairiausias charakteristikas. Straipsnyje [1] yra pasiūlytas modelis, kuris aprašo visų mokslinės srities raktinių žodžių aibės dydį kaip Puasono procesą bei nurodo, kokiu būdu kinta pastebėtų raktinių žodžių kiekis. Tame pačiame straipsnyje įrodyta keletas modelio asimptotinių sąvybių, pasiūlyti parametrų įverčių skaičiavimo būdai. Šiame darbe aprašytas eksperimentinis tyrimas suteikia galimybes teorinių samprotavimų rezultatus panaudoti realiems skaičiavimams.

2. Kai kurie modelio žymėjimai

Nagrinėjama mokslo sritis F ir stebima jos terminijos raida. Pilnas modelis pateiktas [1], čia tik nurodysime keletą žymėjimų ir prielaidų, kurios bus naudojamos darbe:

- pasirodžiusių raktinių žodžių iki laiko momento t aibė – K_t.$X_t = \text{card } K_t$;
- stebėtų atrinktuose žurnaluose raktinių žodžių iki laiko momento t aibė – $S_t (S_t \subset K_t)$.$Y_t = \text{card } S_t$;
- pokyčiai $\Delta X_t = X_t - X_{t-1}$ yra nepriklausomi ir pasiskirstę pagal Puasono dėsnį su parametru λ_t.$N_t = \mathrm{E}X_t$, $\lambda_t = N_t - N_{t-1}$. N_0 – pradinių raktinių žodžių skaičius;
- raktiniai žodžiai sunumeruoti pagal jų pasirodymo momentą ir jiems priskirti svoriai W_k. kurie yra nepriklausomi, vienodai pasiskirstę atsitiktiniai dydžiai. $\mathrm{E}W_k \equiv 1$;
- u_t– tam tikra determinuota funkcija, nusakanti vidutinį stebimų raktinių žodžių kiekį laikotarpiu t;
- dvi laiko eilutės α_t ir $\beta_t$vadinamos asimptotiškai ekvivalenčiomis ($\alpha_t \sim \beta_t$), jei egzistuoja funkcija $f(x)$, kuriai $\lim_{x\to\infty} f(x) = 0$ ir $\mathrm{E}|\alpha_t - \beta_t| \leqslant \mathrm{E}[f(\beta_t)\beta_t]$.

3. Modelio sąvybių analizė Monte-Karlo metodu

3.1. Stebint raktinius žodžius, pasirodžiusius atrinktuose leidiniuose, ar tiksliau – stebėtų terminų kiekį Y_t, norėtųsi turėti išraišką šio dydžio vidurkiui EY_t paskaičiuoti pagal modelio parametrus. Deja, ją gauti ir panaudoti praktikoje yra labai sudėtinga todėl [1] yra pasiūlyta aproksimacija:

$$M_t = N_t - \sum_{\tau=1}^{t} d_\tau E \exp\left\{-W \sum_{l=\tau}^{t} \frac{u_l}{N_l}\right\}, \quad \text{kur } W = W_1 \text{ ir } d_1 + \ldots + d_\tau = N_\tau.$$

Dydžio M_t naudingumas slypi tame, kad esant tam tikroms sąlygoms teisingas sąryšis $M_t \sim Y_t \sim EY_t$ (žr. [1] Theorem 1). Šis sąryšis yra asimptotinis ir nenurodo, kaip greitai ir tiksliai aproksimacija pasireiškia, todėl norint nustatyti jo panaudojimo tikslingumą praktiniuose skaičiavimuose buvo atlikta Monte-Karlo analizė:

- generuojama daug (10, 100, ...) sekų $Y_t^{(i)}$ su vienodais parametrais,
- sekos suvidurkinamos: $\bar{Y}_t = \frac{1}{n} \sum_{i=1}^{n} Y_t^{(i)}$,
- atliekamas $\bar{Y}_t$ ir M_t reikšmės (ji vienoda visiems $Y_t^{(i)}$) palyginimas.

Dydžiams $\bar{Y}_t$ ir M_t palyginti nagrinėjamos tokios charakteristikos:

$$g_t = \frac{\bar{Y}_t}{M_t}; \quad \varphi(\varepsilon) = \min\left\{t: \max_{\tau \geq t} \left|\frac{\bar{Y}_\tau}{M_\tau} - 1\right| < \varepsilon\right\}; \quad B(\upsilon) = \frac{1}{T - \upsilon T} \sum_{i=\upsilon T+1}^{T} \left|\frac{\bar{Y}_i}{M_i} - 1\right|^2.$$

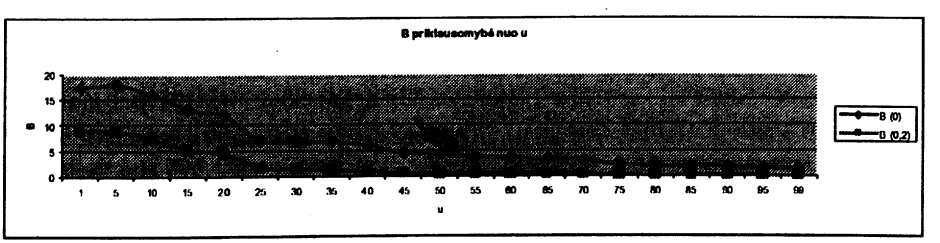
Visos funkcijos (taip pat ir tos, kur bus naudojamos toliau) priklauso nuo daugelio argumentų – λ, u, T... Toliau pateikti šių funkcijų elgesio (didėjimo, mažėjimo, pastovumo), priklausomai nuo atitinkamų argumentų kitimo tyrimai bei jų rezultatai. Kurio konkretaus argumento atžvilgiu analizuojama funkcija bus aišku iš konteksto.

Buvo generuojamos skirtingo ilgio sekos su skirtingais λ ir u (apsiribojama lygių svorių atveju $W_k \equiv 1$) ir nagrinėjamos aukščiau aprašytos charakteristikos su įvairiomis parametrų reikšmėmis ($\varepsilon = 0,1;\ 0,05;\ 0,01;\ \nu = 0;\ 0,2;\ 0,3;\ 0,5;\ 0,7$).

Gauti rezultatai parodė, kad didėjant υ, $B(\upsilon)$ mažėja gana greitai. 1 pav. pavaizduota, kaip tai atrodo sekai, kurios ilgis – 1000, $\lambda = 20, N_0 = 100$.

Iš brėžinio taip pat matyti, kad esant didesnei u reikšmei, B reikšmė mažesnė..

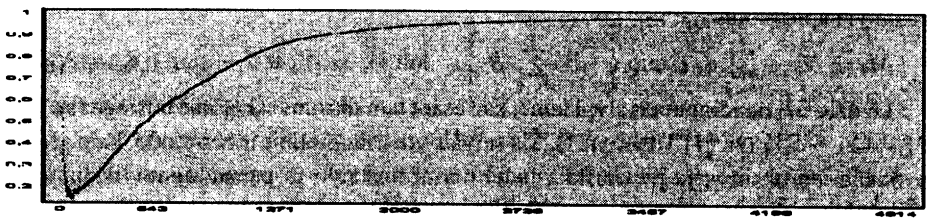
$B(0)$ reikšmės pakankamai didelės, bet užtenka atmesti 200 pirmųjų narių ir gaunamos žymiai mažesnės reikšmės $B(0,2)$. Kai $u = 50$, $B(0,2)$ jau lygus 0,5. Toliau didinant u reikšmes atitinkamai mažėja $B(0,2)$. Jeigu atmestume pusę pirmų narių, tai esant $u = 50$, gautume $B(0,5)$ mažiau už $0,01$.



1 pav. B priklausomybė nuo u.

Realiems duomenims (žr. 5-ąjį skyrelį) situacija labai panaši – didėjant $v$reikšmei, $B(v)$ mažėja. Štai pavyzdžiui $B(0) \approx 0,06$, o $B(0,5) \approx 0,03$. Tiesa, santykis nėra ypač didelis, bet tai galima pagrįsti tuo, kad šiuo atveju seka yra palyginti trumpa.

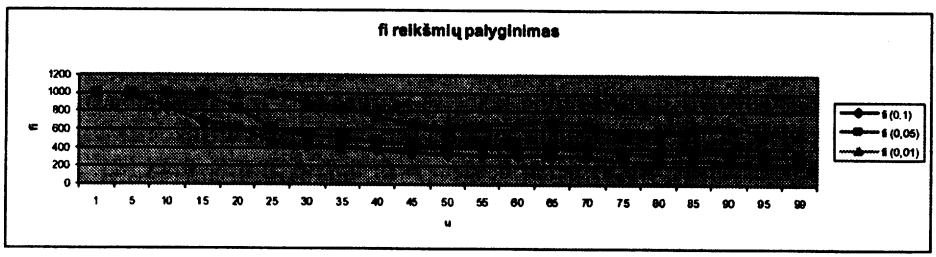
g_t grafikas bendru atveju atrodo taip:



2 pav. g_t grafikas.

Pradžioje yra "duobė", bet labai greitai kreivė išsilygina ir monotoniškai artėja į vienetą. M_t tuo tiksliau įvertina EY_t reikšmes, kuo didesni santykiai $\frac{u_t}{N_t}$. Kai λ pastovus, didesnį u atitinka tikslesnis įvertis. Jei u pastovus, tai didesnį λ atitinka mažiau tikslus įvertis. Esant didesniems santykiams $\frac{u_t}{N_t}$ atitinkamai mažėja "duobė" grafike ir pakilimas link vieneto tampa staigesniu.

Toms pačioms 1000 ilgio sekoms ($\lambda = 20, N_0 = 100$) turime tokius $\varphi(\varepsilon)$ grafikus:



3 pav. φ reikšmių palyginimas.

Matyti, kad u reikšmei didėjant visų $\varphi(\varepsilon)$ reikšmės linkusios mažėti. Visgi, esant pakankami mažoms ε reikšmėms (0,01), turime banguotą kreivę, tačiau tai yra natūralu dėl išliekančio atsitiktinumo, kuris pasireiškia labai nežymiai, tačiau pakankamai, kad būtų pastebimas.

Analizės rezultatai perša išvadą, kad M_t aproksimuoja EY_t labai gerai, kai turime sekas, kurių ilgis maždaug apie 1000 ir daugiau. Net ir šiuo atveju reikėtų prisiminti, kad pradžioje esama "duobės" – prasto atitikimo, bet atmetus nedidelį kiekį pradinių narių (iki 20 – 30% priklausomai nuo $\frac{u_t}{N_t}$ reikšmių), sutapimas žymiai pagerėja ir greitai pasiekia 90 – 100%. Jei turime sekas, kurių ilgis yra mažesnis – keli šimtai, tenka būti atsargesniems, nes galimi variantai, kai sutapimas neperlipa nė 70% ribos. Visgi realūs duomenys (žr. 5 skyrelį) parodė, kad net esant labai trumpai sekai (200) M_t reikšmės labai tiksliai įvertina Y_t vidurkio reikšmes.

3.2. Žemiau pateikiama keletas pastabų ir patarimų dėl M_t skaičiavimo praktikoje. Šiame darbe buvo apsistota ties dviem atvejais: a) $W_k \equiv 1$ ir b) W_k pasiskirstę pagal Pareto dėsnį:

$\forall x > c_0\ \mathrm{P}\{W_k > x\} = \left(\frac{c_0}{x}\right)^\gamma, \gamma > 1$, o $c_0 = c_0(\gamma)$ užtikrina sąlygą $\mathrm{E}W_k \equiv 1$.

Pirmuoju atveju M_t skaičiavimas akivaizdus – tiesiog nelieka matematinio vidurkio. Antruoju atveju turime sudėtingesnį variantą. Kadangi Pareto skirstinio vidurkis yra $c_0\frac{\gamma}{\gamma-1}$, tai turime, kad $c_0 = \frac{\gamma-1}{\gamma}$. Pažymėkime $\sum\limits_{l=\tau}^{t} \frac{u_l}{N_l} = \theta(\tau)$. Tada $\mathrm{E}\exp\{-W\theta(\tau)\} = \int\limits_{c_0}^{+\infty} \exp\{-\theta(\tau)x\} d\left(1 - \frac{c_0^\gamma}{x^\gamma}\right) = \gamma c_0^\gamma(\theta(\tau))^\gamma \int\limits_{c_0\theta(\tau)}^{+\infty} \frac{e^{-x}}{x^{\gamma+1}}\mathrm{d}x$. Pažymėkime $H\gamma(a) = \int\limits_a^{+\infty} \frac{e^{-x}}{x^\gamma}\mathrm{d}x$. $G\gamma(a) = \sum\limits_{i=0}^{+\infty} (-1)^i \frac{C_i}{i!}$, kur $C_i = \begin{cases} \frac{a^{i-\gamma+1}}{i-\gamma+1}, & i \neq \gamma - 1 \\ \ln a, & i = \gamma - 1 \end{cases}$. Nesunku pastebėti, kad $H\gamma(a) = G\gamma(\infty) - G\gamma(a)$, $G\gamma(\infty) = \lim_{x\to\infty} G\gamma(x)$. Pažymėkime $H^*\gamma(a) = -G\gamma(a)$. Tada $H\gamma(a) - H^*\gamma(a) = G\gamma(\infty)$. Iš čia gauname: $H\gamma(1) - H^*\gamma(1) = G\gamma(\infty)$. Taigi $H\gamma(a) = H^*\gamma(a) + (H\gamma(1) - H^*\gamma(1))$.

Praktiniams skaičiavimams siūlomas toks kelias:

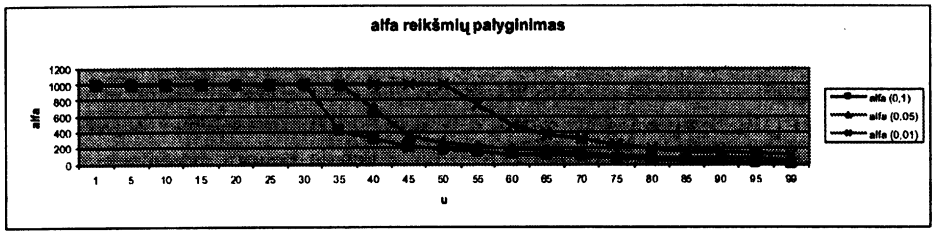
- iki norimo tikslumo skaičiuojama $H^*\gamma(a)$ reikšmė;
- pridedama prie rezultato $(H\gamma(1) - H^*\gamma(1))$ reikšmė, kurią galima paskaičiuoti kokiu nors matematiniu paketu (Maple). Galima netgi turėti iš anksto pasiruoštą lentelę su tomis reikšmėmis įvairiems γ.

3.3. Dar viena nagrinėjama modelio sąvybė – kaip kinta kiekis raktinių žodžių, kurie buvo pradiniu (nuliniu) laiko momentu ir dar nepastebėti iki laiko momento t. Theorem 3 [1] teigia, kad $\lim_{t\to\infty} \mathrm{P}\{K_0 \in S_t\} = 1 \iff \sum_{t=1}^{\infty} \frac{u_t}{N_t} = \infty$. Tačiau ši teorema nieko nepasako apie tai, kaip ši charakteristika elgiasi, esant baigtinėms t reikšmėms, taip pat neaišku, nuo kurio laiko momento $K_0 \in S_t$. Siekiant visa tai ištirti buvo nagrinėjamas dydis $D_t = \frac{card(K_0 \setminus S_t)}{card K_0}$, kuris parodo, kokia dalis pradinių raktinių žodžių dar liko nepastebėta. Tiriant dydį D_t buvo įsivesta dar viena funkcija, kuri labai panaši į $\varphi(\varepsilon)$: $\alpha(\varepsilon) = \min\{t\colon D_t < \varepsilon\}$. Ji parodo, kaip greitai D_t reikšmė tampa kiek norima artima nuliui.

4 pav. pateiktas grafikas parodo, kaip elgiasi $\alpha(\varepsilon)$ reikšmės, kai didėja u (T = 1000, $\lambda = 20, N_0 = 100$). Matome, kad pradedant nuo $u = 55$, $\alpha(0,01) < 1001$, t.y, esama momento, nuo kurio pradedant lieka nepastebėta nedaugiau, kaip 1% raktinių žodžių. Kuo toliau didiname u reikšmė, tuo greičiau užfiksuojami 99% pradinių raktinių žodžių. Kadangi S_t reikšmė nepriklauso nuo to, kokio ilgio sekos gabalas lieka už momento t iki galo, tai galima padaryti įvairių išvadų: sekoms, kurių ilgis 500, o $u < 60\alpha(0,01)$ yra neapibrėžta, t.y., nėra tokios t reikšmės, nuo kurios pradedant visi D_t būtų nedidesni už 0,01.

Didinant sekos ilgį vis mažesniems ε $\alpha(\varepsilon)$ įgyja apibrėžtą reikšmę.

Rezultatai patvirtino, kad D_t reikšmė tikrai mažėja ar net artėja į nulį. Šio nykimo greitis priklauso nuo $\frac{u_t}{N_t}$ dydžių. Jei N_t auga labai staigiai, lyginant su u_t, t.y., yra aprėžta iš viršaus arba bent jau $\sum\limits_{t=1}^{T} \frac{u_t}{N_t}$ yra labai mažas (tarkim, $\frac{1}{1000}T$), tai D_t niekada nepasiekia pakankamai mažos reikšmės arba ją pasiekia po labai ilgo laiko tarpo. Tuo atveju,



4 pav. *alpha* reikšmių palyginimas.

kai $\sum_{t=1}^{\infty} \frac{u_t}{N_t} = \infty$, tai tikrai žinome, kad $D_t \to 0$. Jei u – pastovus (juo labiau augantis dydis), $\frac{u}{N_0}$ pakankamai skiriasi nuo nulio (bent jau $\frac{1}{10}$ eilės) bei $N_t - N_{t-1}$ yra nedidelis dydis, lyginant su N_t (ar N_0), tai šis konvergavimas pasireiškia greitai – užtenka kokių 500 – 1000 ilgio sekos, kad būtų pastebėta ne mažiau kaip 80 – 90% pradiniu laiko momentu buvusių raktinių žodžių. Aukščiau esama keletos neapibrėžtų didumo ar mažumo sąvokų, bet jos priklauso nuo konkrečios situacijos. 5-ame skyrelyje panagrinėtas konkretus diskretinės matematikos terminijos sukauptos duomenų bazės atvejis (u – pastovus, $\frac{u}{N_0} \approx \frac{1}{20}$, o $N_t - N_{t-1} = \lambda \approx \frac{1}{20} N_0$), kuriame 200 ilgio sekos užtenka, kad būtų pastebėta 92% pradinių raktinių žodžių.

4. Modelio parametrų įverčiai ir jų tikslumo analizė

4.1. Vienas iš svarbiausių modelio parametrų – dydis λ, kuris charakterizuoja mokslinės srities augimo tempus. Šio dydžio įverčiams skaičiuoti [1] buvo pasiūlyti du būdai. Dėl straipsnio apimties apribojimo čia detaliau panagrinėtas paprastesnis variantas. Apsiribosime atvejais, kai $\lambda_t \equiv \lambda$ ir $u_t \equiv u$ bei $W_k \equiv 1$.

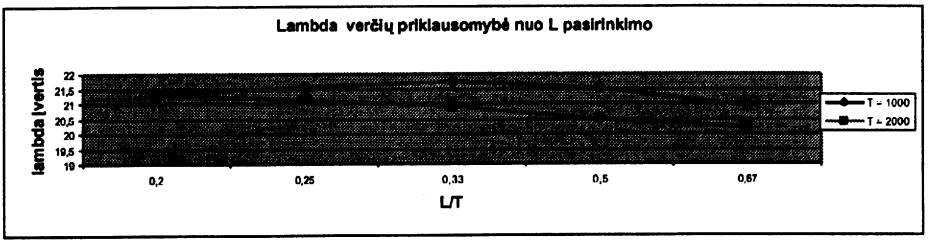
Pažymėkime $\widehat{a} = \frac{Y_T - Y_L}{T-L}$, kur $L = L(T)$ tenkina sąlygą $C_2 T \leqslant L \leqslant C_1 T$, čia $0 < C_2 < C_1 < 1$. Tokia išraiška naudojama ne atsitiktinai – tai yra [1] aprašyto dydžio a, lygaus ribai $\lim_{t\to\infty} \mathrm{E}\Delta Y_t$, įvertis. Tada λ įvertis: $\widehat{\lambda} = \frac{\widehat{a}u}{u-\widehat{a}}$. Vietoje u galima imti jos įvertį: $\widehat{u} = \frac{1}{T} \sum_{i=1}^{T} cardQ_i$. Taigi turime, kad $\widehat{\lambda} = f(L)$. Dominantis klausimas – koks λ įverčio $\widehat{\lambda}$ tikslumas ir kaip priklauso nuo pasirinkto L.

Buvo generuojamos įvairaus ilgio sekos su skirtingomis λ reikšmėmis (1, 5, 10, 20, 50) bei skirtingomis u_t reikšmėmis (1, 10, 25, 50, 75, ...). Po to gautoms sekoms buvo skaičiuojamas $\widehat{\lambda}$ įvairioms L reikšmėms ($\frac{1}{5}T, \frac{1}{4}T, \frac{1}{3}T, \frac{1}{2}T, \frac{2}{3}T$).

Rezultatai patvirtino akivaizdžią tiesą: skirtingo ilgio sekoms $\widehat{\lambda}$ įvertis gaunamas nevienodo tikslumo, bet vidutiniškai neatitikimas yra linkęs mažėti ilginant sekas. 5 pav. matyti, kad kai λ tiksli reikšmė yra 20 ir turime dvi skirtingo ilgio sekas (1000 ir 2000), ilgesniosios sekos įverčiai tikslesni visoms L reikšmėms.

Taip pat esama priklausomybės nuo u – didesnėms reikšmėms įvertis tikslesnis. Tiksliausi įverčiai gaunami, kai u reikšmė yra lygi pusei N_0 arba bent jau panašios eilės.

Tačiau svarbiausias dominantis klausimas – kaip įvertis priklauso nuo L reikšmės.



5 pav. λ verčių priklausomybė nuo L pasirinkimo.

Išnagrinėjus visus rezultatus paaiškėjo, kad įverčiai mažiausiai ir didžiausiai L reikšmei ($\frac{1}{5}T$ ir $\frac{2}{3}T$) gali skirtis pakankamai daug – net iki 25%, kai turimos labai neilgos sekos ir mažos u reikšmės. Tačiau jei didiname u reikšmę, tai skirtumas labai greitai mažėja. Atveju, kai $N_0 = 100$, o $u = 25$ skirtumas nesudaro nė 10%. O jeigu $u = 50$, tai skirtumas tarp mažiausios ir didžiausios įverčio reikšmės tampa praktiškai nepastebimas – iki 5%. Bendru atveju įverčio tikslumas linkęs didėti, kai didėja L reikšmė, kaip tai matyti 5 pav. grafike aukščiau (■ linija), bet galimas ir toks variantas, kai tame pačiame grafike matoma ♦ linija – tikslumas pereinant nuo vienos L reikšmės prie didesnės gali blogėti, bet labai nežymiai. Kuo didesnė λ reikšmė, tuo aiškesnė įverčio gerėjimo tendencija.

Tiksliausia įverčio reikšmė skiriasi nuo tikslios parametro reikšmės pakankamai nedaug – keliais procentais į vieną ar kitą pusę. Visoms generuotoms sekoms, kurių u buvo lygus pusei N_0 įvertis, kai $L = \frac{2}{3}T$(nebūtinai tiksliausias) nuo tikslios reikšmės vidutiniškai skyrėsi maždaug 4%. Tiksliausioms įverčio reikšmėms šis skaičius sumažėja maždaug iki 3 – 3,5%.

4.2. Straipsnyje [1] pasiūlytas kitas λ įverčio skaičiavimo būdas remiasi maksimalaus tikėtinumo metodologija, tik yra šiek tiek supaprastintas. Kaip jau minėta aukščiau, šio įverčio tikslumas nenagrinėjamas š iame darbe, bet pats sąryšis pasitarnauja kitam tikslui – N_0 įvertinimui realių duomenų atveju.

Turime $N_t = n + \widehat{\lambda}t$, čia $\widehat{\lambda}$– įvertis, aprašytas 4.1 skyrelyje, o n – nežinomas dydis (lygus N_0). Jo įvertį galima gauti naudojantis [1] (20) lygybe:

$$\sum_{t=2}^{T} \frac{\Delta Y_t \cdot Y_{t-1} - \zeta_t (N_t(\theta) - Y_{t-1})}{N_t^2(\theta)} N_t'(\theta) = 0, \quad \text{kur } \theta = (n, \lambda) \text{ ir } \zeta_t = \text{card } Q_t \cap S_{t-1}.$$

Mes turime: $\frac{dN_t}{dn} = 1$, vietoje λ statome $\widehat{\lambda}$ ir turime tokią lygybę:

$$\sum_{t=2}^{T} \frac{\Delta Y_t \cdot Y_{t-1} - \zeta_t (n + \lambda t - Y_{t-1})}{(n + \lambda t)^2} = 0. \qquad (*)$$

Toliau išskyrę (*) kairiąja pusę į dvi funkcijas $A_n = \sum_{t=2}^{T} \frac{\Delta Y_t \cdot Y_{t-1} - \zeta_t (\lambda t - Y_{t-1})}{(n+\lambda t)^2}$ ir $B_n = \sum_{t=2}^{T} \frac{\zeta_t n}{(n+\lambda t)^2}$ galime ieškoti, kur jos pakeičia savo skirtumo ženklą.

5. Realūs duomenys (DISC)

Iš "Elsevier" leidyklos ir firmos VTEX buvo gauta žurnalo "Discrete Mathematics" raktinių žodžių duomenų bazė. Po pertvarkymų sudaryta Y_t laiko eilutė, kurios ilgis 200 (laiko vienetas realiame gyvenime atitinka mėnesį). Šiai eilutei apskaičiuoti λ, u ir N_0 įverčiai. Vertinant λ, L buvo parinktos reikšmės 0, 10, 20, ..., 190. Žemiau pateikti įverčiai, apskaičiuoti $50 \leqslant L \leqslant 150$ reikšmėms:

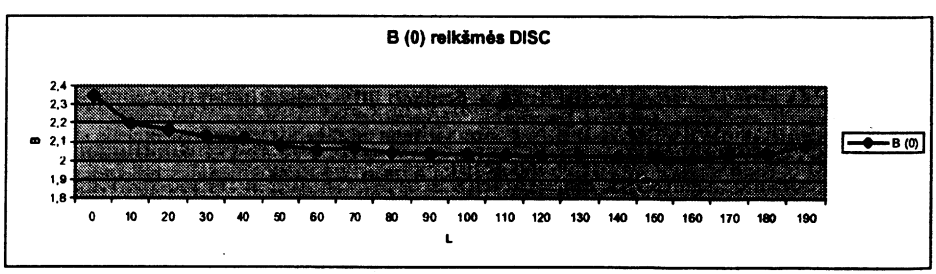
Gauti įverčiai buvo panaudoti M_t skaičiavimui. Po to Y_t ir M_t buvo palyginti, naudojant charakteristiką $B(v)$. 6 ir 7 pav. pateikti $B(0)$ ir $B(0,2)$ grafikai.

Iš jų matyti, kad negalima tiksliai nustatyti, kuris įvertis geresnis. Jei žiūrėtume į visą M_t ir Y_t sutapimą ($B(0)$), tai geriausias λ įvertis būtų, kai $L = 160$, t.y. $\approx \frac{8}{10}T$, bet jei atmesime pradžioje 40 narių ($B(0,2)$), tai gausime, kad geresnis įvertis yra pirmojoje pusėje. Beje, ir kiti $B(v)$, kur $v > 0,2$ blogėja L artėjant prie T, t.y., kreivės užriestos į viršų. Jeigu pažiūrėtume į M_t ir Y_t grafikus, tai matytume, kad pirmuoju atveju esama geresnio sutapimo pradžioje, bet gale kreivės išsiskiria. Antruoju gi atveju, kreivės praktiškai sutampa visur, išskyrus nedidelę dalį pradžioje.

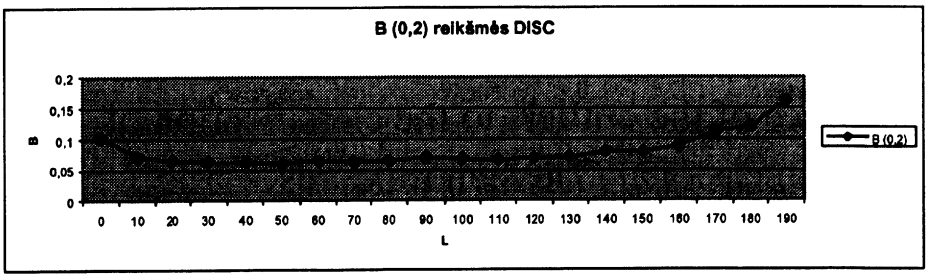
D_t reikšmė mažėja iki maždaug 0,08, t.y., galima teigti, kad per stebėjimo laiką buvo pastebėta apie 92% raktinių žodžių, buvusių pradiniu laiko momentu.

1 lentelė.

Nr.	L	lambda	Ut	N0
1	60	209	240	4982
2	70	212	240	4833
3	80	205	240	5211
4	90	201	240	5419
5	100	202	240	5367
6	110	201	240	5447
7	120	199	240	5550
8	130	199	240	5541
9	140	190	240	6076
10	150	190	240	6080



6 pav. $B(0)$ grafikas.



7 pav. $B(0, 2)$ grafikas.

6. Išvados

Darbo metu atlikta analizė parodė, kad asimptotinės modelio sąvybės pasireiškia pakankamai greitai, neblogas tikslumas pasiekiamas jau po kelių šimtų stebėjimų. Realūs duomenys ir sveikas protas sufleruoja, kad tikros sekos panašaus ilgio ir bus – jų ilgiai sieks kelis šimtus ar dar net mažiau. Net ir tokiais atvejais esant išpildytoms kai kurioms sąlygoms galima kuo puikiausiai pasitikėti priartėjimais ir įverčiais. Kitu atveju neišvengiamos nemenkos paklaidos ar net visiškai priešingi rezultatai, nei tikimasi. Visgi tikėtina, kad realūs duomenys bus palankūs, kaip rodo konkretus išnagrinėtas diskretinės matematikos atvejis.

Literatūra

[1] M. Hazewinkel, R. Rudzkis, Probabilistic model for the growth of Thesauri, *Acta Applicandae Mathematicae*, **00**, 1–16 (2001).

[2] С.А. Айвазян, В.С. Мхитарян, *Прикладная статисика и основы эконометрики*, Юнити, Москва (1998).

Experimental analysis of models for thesauri

V. Balys

Estimating and comparing the rate of growth of various science fields becomes very interesting and essential problem. Evolving sets of keyphrases (and keywords) observed in representative journals carry all needed information about this development. A model for the growth of thesauri as well as asymptotic properties and estimators for the parameters are proposed in [1]. The results of experimental analysis enable these properties and estimators to be used for practical calculations.

The analysis suggests that sufficient accuracy of approximations is achieved quite fast – for sequences of few hundred length.

Statistical analysis of "Discrete Mathematics" journal keyphrases database illustrate the use and the reliability of estimations.