Liet. matem. rink., T. 43, spec. nr., 463–467



# Mokslinių terminų statistinio pasiskirstymo taikymas straipsnių klasifikavime

Vaidas BALYS, Rimas RUDZKIS (MII)
*el. paštas:* rudzkis@ktl.mii.lt

## Įvadas

Dideliais tempais augant elektroninės leidybos mastams bei plečiantis Interneto tinklui elektroninės mokslinės informacijos bibliotekos tampa neišvengiama būtinybe. Tokiose bibliotekose bene svarbiausia yra lanksčios ir naudotojui patogios paieškos priemonės, kurios galimos tik išsprendus informacijos atpažinimo bei klasifikavimo uždavinius bei įdiegus konkrečius sprendimus. Pastarųjų uždavinių kontekste itin aktualia bei perspektyvia idėja tampa automatizuotas straipsnių indeksavimas (tuo pačiu ir klasifikavimas).

Vienas iš galimų publikacijų klasifikavimo būdų – automatizuotai priskirti joms tam tikrą kiekį mokslinių terminų iš kontroliuojamo sąrašo. Autoriaus išskirti raktiniai žodžiai čia nėra tinkami dėl keleto priežasčių: dėl skirtingu metu ar/ir skirtinguose leidiniuose buvusių nevienodų reikalavimų, arba dėl autorių subjektyvumo skiriasi raktinių žodžių kiekis publikacijose (senesniuose numeriuose jų dažnai net nėra), taip pat skiriasi jų parinkimo principas (raktiniu žodžiu gali būti laikoma tiek naujai įvesta sąvoka, tiek konkretus dažnai tekste sutinkamas terminas, tiek bendros sąvokos). Kita svarbi priežastis yra nelankstumas – autorius išskiria raktinius žodžius vieną kartą pagal tam tikrus reikalavimus, o sprendžiant įvairius taikomuosius uždavinius gali prireikti kito skaičiaus ar kitokių raktinių žodžių.

Automatizuotas raktinių žodžių priskyrimas straipsniams gali būti taikomas daugelyje praktinių uždavinių: mokslo srities ar mokslo žurnalo plėtros įvertinimas [1]; raktinių žodžių priskyrimas straipsniams pagal leidinio reikalavimus; atskirų leidybos objektų ar kitų veiklos sričių atitikimų nustatymas; paieškos sistemos; mokslinių straipsnių klasifikavimas standartinėse klasifikavimo sistemose ir t.t.

## Žymėjimai

$A$ – straipsnių aibė;
$T(a)$ – straipsnio $a \in A$ tekste esantys mokslo terminai;
$W(a)$ – straipsnio $a \in A$ žodžiai. Čia ir toliau „žodžiais“ vadinsime tiek žodžius, tiek jų junginius (frazes);
$K(a)$ – straipsnio $a \in A$ raktiniai žodžiai, $\widehat{K(a)}$ – statistinis įvertis;
$f_w(a)$ – žodžio $w$ pasikartojimų skaičius straipsnyje $a \in A$; $f_w^l(a, v)$ – skaičius $a \in A$ sakinių, kuriuose yra $v$ ir $w$ bei atstumas tarp jų neviršija $l$;

$G(a)$ – straipsnio $a \in A$ tekste esančios specialios kalbos frazės, ypač dažnai sutinkamos su raktiniais žodžiais įvairiuose mokslo leidiniuose.

Čia ir toliau kiekvienai aibei $B$ jos elementų skaičių žymėsime $|B|$.

## Automatizuoto raktinių žodžių priskyrimo straipsniams algoritmai

Formaliai raktinių žodžių priskyrimo moksliniam straipsniui uždavinį būtų galima suformuluoti taip: turime straipsnį $a \in A$, reikia suformuoti $\widehat{K(a)}$. Toliau pateiksime keletą algoritmų, sprendžiančių šį uždavinį.

*Algoritmų klaidų matai*

Siekdami įvertinti bei palyginti $\widehat{K(a)}$ skaičiavimo algoritmus, apibrėžiame dviejų tipų klaidas: pirmojo tipo klaida – autoriaus išskirtas raktinis žodis nėra tarp algoritmo surastųjų; antrojo tipo klaida – algoritmo rastas raktinis žodis nėra tarp autoriaus išskirtųjų. Tada algoritmų efektyvumo matą galima apibrėžti kaip atitinkamus jo daromų klaidų dalies įverčius.

$$\alpha = \frac{1}{|A|} \sum_{a \in A} \frac{|K(a) \backslash \widehat{K(a)}|}{|K(a)|}, \tag{1}$$

$$\beta = \frac{1}{|A|} \sum_{a \in A} \frac{|\widehat{K(a)} \backslash K(a)|}{|\widehat{K}(a)|}. \tag{2}$$

$\alpha$ ir $\beta$ vadinsime atitinkamai pirmo tipo klaidų bei antro tipo klaidų kiekiu. Galima apibrėžti vieną bendrą klaidos matą kaip tam tikrą pirmo ir antro tipo klaidų kombinaciją. Mes naudosime tiesinį darinį

$$\gamma(\gamma_\alpha, \gamma_\beta) = \frac{\gamma_\alpha * \alpha + \gamma_\beta * \beta}{\gamma_\alpha + \gamma_\beta}. \tag{3}$$

*Pastaba* 1. $\gamma$ naudoti kaip minimizavimo kriterijų reikia labai atsargiai, ypač tuo atveju kai vienas iš svorių $\gamma_\alpha$ ir $\gamma_\beta$ lygus 0.

*Pastaba* 2. Klaidų $\alpha$ ir $\beta$ apibrėžimuose yra trūkumas – klaida vertinama pagal autoriaus išskirtus raktinius žodžius, tad atsižvelgiant į įvade pateiktus samprotavimus akivaizdu, jog šie matai nėra visiškai objektyvūs ir neturi prasmės tuo atveju, jei straipsnis neturi priskirtų raktinių žodžių.

*Pastaba* 3. Klaidų įverčius $\alpha$ $\beta$ būtų teisingiau apibrėžti naudojant ne aibes $K(a)$ ir $\widehat{K(a)}$, bet jų plėtinius $K^*(a)$ ir $\widehat{K(a)}^*$, sudarytus šiuo būdu:

$$K^*(a) = K_0^*(a) \bigcup K_1^*(a), \tag{4}$$

kur $K_0^*(a)$ yra visi $K(a)$ žodžiai ir junginiai bei visi jų sinonimai, o $K_1^*(a)$ yra visi $K(a)$ žodžių ir junginių pirmo laipsnio apibendrinimai bei šių apibendrinimų sinonimai. $\widehat{K(a)}^*$ gaunamas kaip plėtinys $(\widehat{K(a)})^*$.

*Mokslinių terminų pasiskirstymo dėsniai*

Tikėtina, kad iškeltojo uždavinio sprendinys bus kelių skirtingus aspektus užčiuopiančių algoritmų kombinacija, todėl natūralu pradėti nuo pačių paprasčiausių modelių ir patikrinti jų panaudojimo galimybes. Beveik akivaizdu, kad šie algoritmai neduos itin gerų rezultatų, tačiau jų tyrimas parodys kokia linkme reikėtų ieškoti tobulesnių sprendimų.

*Dažnumų skaičiavimas.* Viena paprasčiausių ir daug kur taikomų idėjų paremta dažnuminių terminų charakteristikų skaičiavimu. Esant prielaidai, kad straipsnis $a \in A$ parinktas atsitiktinai, pažymėkime

$$P_w(k) = P\{f_w(a) \geqslant k\}, \tag{5}$$

$$p_w(a) = P_w(f_w(a)), \tag{6}$$

$$\pi_w(\delta) = P\{w \in K(a) | p_w(a) \leqslant \delta\}, \quad \delta \in [0..1]. \tag{7}$$

Tada algoritmas atrodytų taip:

$$\widehat{K(a)}_{(\Theta)} = \bigcup_{w \in W(a)} \{w\colon\ \pi_w(p_w(a)) \leqslant \theta_w, \theta_w \in \Theta\}, \tag{8}$$

čia $\theta_w \in [0..1]$ – slenksčio parametrai kiekvienam žodžiui.

*Sąryšiai su specialiaisiais kalbos žodžiais*

Šio algoritmo idėja paremta hipoteze, kad kalboje esama tam tikrų nusistovėjusių frazių, kurios signalizuoja, kad tame sakinyje kalbama apie labai svarbius dalykus, t.y., šalia tų frazių su didele tikimybe sutinkamas straipsnio raktinis žodis. Tegul

$$\pi_w(k,l) = P\left\{w \in K(a) | \sum_{v \in G(a)} f_w^l(a,v) \geqslant k\right\}, \tag{9}$$

tada algoritmas apibrėžiamas analogiškai kaip (9):

$$\widehat{K(a)}_{(\Theta,k,l)} = \bigcup_{w \in W(a)} \{w\colon\ \pi_w(k,l) \leqslant \theta_w, \theta_w \in \Theta\}. \tag{10}$$

*Kombinuotas algoritmas*

Dar viena idėja – apjungti aukščiau aprašytus algoritmus į vieną sudėtingesnį tokiu būdu:

$$\widehat{K(a)}_{(\Theta^1,\Theta^2,\Theta^3,k,l)} =$$

$$= \bigcup_{w \in W(a)} \left\{ w\colon \pi_w(p_w(a)) \leqslant \theta_w^1 \bigvee (\pi_w(p_w(a)) \leqslant \theta_w^2 \bigwedge \pi_w(k,l) \leqslant \theta_w^3) \right\}, \quad (11)$$

čia $\theta_w^1 \in \Theta^1$, $\theta_w^2 \in \Theta^2$, $\theta_w^3 \in \Theta^3$ – slenksčių parametrai, $\theta_w^1 < \theta_w^2$, o ženklai $\bigvee$ ir $\bigwedge$ žymi logines „arba“ bei „ir“ operacijas.

## Tyrimo rezultatai, išvados bei pasiūlymai

Praktinis pasiūlytų algoritmų tyrimas buvo atliktas naudojantis keleto tikimybių teorijos ir statistikos sričių žurnalų straipsnių duomenų baze (iš viso daugiau nei 300 straipsnių). Palyginimui pateikti dar vieno reliai taikomo praktikoje algoritmo, pagrįsto lingvistiniais principais, rezultatai. Lentelėje pateikiami bendrosios klaidos minimumai įvairiems klaidų svoriams.

**Bendros klaidos įverčiai**

| *Svoriai* | *Dažnumų skaičiavimas* | *Sąryšiai su spec. žodžiais* | *Lingvistinis* |
|---|---|---|---|
| $\gamma_\alpha = 1,\ \gamma_\beta = 0$ | 0.3706 | 0.6338 | 0.4930 |
| $\gamma_\alpha = 1,\ \gamma_\beta = 1$ | 0.6020 | 0.7808 | 0.6330 |
| $\gamma_\alpha = 2,\ \gamma_\beta = 1$ | 0.5622 | 0.7318 | 0.5863 |
| $\gamma_\alpha = 3,\ \gamma_\beta = 1$ | 0.5163 | 0.7073 | 0.5630 |

Rezultatai patvirtino įtarimus, kad paprasčiausi algoritmai neduos itin gerų rezultatų, nors matyti, kad paprastas dažnumų skaičiavimas veikia taip pat arba net truputį efektyviau, nei praktikoje naudojamas lingvistinis algoritmas. Lentelėje nepateikti duomenys apie kombinuoto algoritmo rezultatus dėl tam tikrų techninių sunkumų suskaičiuoti minimumą, bet bendru atveju jis labai nežymiai (keliomis šimtosiomis) pagerina pirmąjį algoritmą.

Greta šio formalaus tyrimo buvo atlikta ekspertinė analizė, siekiant išsiaiškinti du dalykus: kiek yra netikrų klaidų, t.y., tam tikrų situacijų, kai algoritmas fiksuoja klaidą, o žmogus jos nefiksuotų (pvz. sinonimai), bei kiek ir kokiu būdu būtų galima sumažinti klaidas, t.y., kuria linkme galima tobulinti algoritmus. Šiam tyrimui atsitiktiniu būdu buvo atrinki 5% straipsnių. Gauti rezultatai (vidurkiai) dažnumų skaičiavimų algoritmui pateikti žemiau esančioje lentelėje:

| $\lvert K(a)\rvert$ | $\lvert\widehat{K(a)}\rvert$ | $\lvert K(a) \bigcap \widehat{K(a)}\rvert$ | $\lvert D1\rvert$ | $\lvert D2\rvert$ | $\lvert D3\rvert$ | $\alpha$ | $\alpha^*$ |
|---|---|---|---|---|---|---|---|
| 5 | 17.36 | 2.36 | 3.43 | 1.64 | 0.86 | 0.4815 | 0.2362 |

Čia $D1$, $D2$ ir $D3$ apibrėžiami kaip aibės tų terminų iš $\widehat{K(a)}$, kurie yra sudėtinės dalys kitų žodžių, esančių $\widehat{K(a)}$ ($D1$) arba $K(a) \bigcap \widehat{K(a)}$ ($D2$), arba $K(a) \setminus \widehat{K(a)}$ ($D3$). Todėl antro tipo klaidą būtų galima sumažinti $\widehat{K(a)}$ pakeitus į $\widehat{K(a)} \setminus (D1 \bigcup D2)$, tačiau sumažėjimas nėra itin reikšmingas. Tuo tarpu jei iš $D3$ žodžių sugebėtume atkurti pilnus terminus (papildant kitais žodžiais), esančius aibėje $K(a)$, tai gautume pirmo tipo klaidą žymiai mažesnę (žr. $\alpha^*$ lentelėje). Deja terminą tiesiogiai atkurti yra ypač sudėtinga, nes dažniausiai ne visos jo sudedamosios dalys yra aibėje $D3$.

Išdėstyti algoritmai turi labai didelį trūkumą – jie raktinius žodžius parenka tik iš terminų, esančių tekste, o aukščiau pateiktam pavyzdžiui vidutiniškai 2.43 termino iš 5 raktinių žodžių tiesiogiai tekste nesutinkami. Šia prasme galima teigti, kad dažnumų skaičiavimo algoritmas daro labai mažą pirmo tipo klaidą, nes jis suranda 2.36 iš 2.57 tekste esančių raktinių žodžių. Deja to nepakanka ir reikalingi algoritmai, kurie sugebėtų atkurti bent dalį iš tų 2.43 terminų, tiesiogiai nesančių tekste. Ekspertinis tyrimas taip pat parodė, kad sinonimų, santrumpų ir apibendrinimų atpažinimas leistų papildyti raktinių žodžių sąrašą vidutiniškai ne daugiau kaip 0.3–0.4 nauju terminu, tad sprendimo reikia ieškoti kitur.

Viena galimų ir šiuo metu sparčiai populiarėjančių idėjų – pasinaudoti terminų kontekstine aplinka, kurią būtų galima apibrėžti taip: mokslinis terminas turi tam tikrą aplinką (grupę terminų), kuri dažnai sutinkama tekstuose greta šio termino. Šios aplinkos ar jos dalies buvimas tekste ženkliai padidina tikimybę, kad kalbama būtent apie tą terminą. Idėjos patrauklumas tas, kad pats terminas tiesiogiai gali ir nebūti tekste, arba gali būti tam tikra jo dalis, o gal ir visos dalys, tačiau išsibarsčiusios toli viena nuo kitos. Daugiau apie šią idėją galima pasiskaityti [2, 3]. Pagrindinis uždavinys, kurį dar reikia išspręsti – adekvataus terminų aplinkos modelio sukūrimas. Būtent šio uždavinio sprendimas (aktyviai bendradarbiaujant su pagrindiniu išdėstytosios idėjos iniciatoriumi prof. M. Hazewinkeliu) yra dabartinis mūsu prioritetas, o pirmuosius samprotavimus, idėjas bei rezultatus tikimės pateikti artimiausiose publikacijose.

## Literatūra

[1] M. Hazewinkel, R. Rudzkis, A probabilistic model for the growth of thesauri, *Acta Appl. Math.*, **67**, 237–252 (2001).

[2] M. Hazewinkel, Enriched thesauri and their uses in information storage and retrieval, in: *Proceedings of the First DELOS Workshop*, C. Thanos (Ed.), INRIA, Sophia Antipolis (1997), pp. 27–32.

[3] M. Hazewinkel, Topologies and metrics of information spaces, *CWI Quaterly*, **12**(2), 93–110 (1999).

## Statistical distributions of thesauri and their uses in classification of science publications

V. Balys, R. Rudzkis

Automatic keywords assignment is a way to solve the problem of classification of scientific information. Paper deals with idea of implementing this assignement by making use of models for distribution of scientific terms. Some analysis was carried out and on base of its results the conclusions about effectiveness, usefullnes, main shortages and the ways of refinement of simpliest algorithms are drawn. A need for better algorithms is motivated and some ideas about how to create them are proposed.