Liet. matem. rink., **44**, *spec. nr.*, 2004, 537–541

Mokslo terminų aplinkų modelių taikymas straipsnių klasifikavime

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

Reziumė. Straipsnio raktinių žodžių nustatymas yra tam tikra prasme ekvivalentus klasifikavimui, ir tam galima panaudoti tekstų kategorizavimo metodus. Mokslo informacijos specifika sąlygoja ir kitokių metodų poreikį, todėl šiame darbe pristatoma mokslo terminų aplinkų idėja, o jų matematinis apibrėžimas pateikiamas tikimybinių skirstinių sąvokomis. Pateikiami šių aplinkų panaudojimo raktinių žodžių radimui pasiūlymai.

Raktiniai žodžiai: publikacijų klasifikavimas, raktiniai žodžiai, mokslo terminų aplinkos, statistiniai terminų pasiskirstymo modeliai.

Įvadas

Remiantis kai kuriais šaltiniais, matematinių žinių kiekis padvigubėja kas 10–15 metų, o iki šių dienų išleistus unikalius matematinius darbus pavyktų sutalpinti tik į dešimtimis kilometrų matuojamas lentynas. Panaši situacija yra ir daugelyje kitų mokslo sričių, ir galima drąsiai teigti, kad mokslo žinių apimtys didesniu ar mažesniu tempu augs ir artimiausioje ateityje. Akivaizdu, kad tokiems informacijos kiekiams valdyti reikalingos itin efektyvios priemonės, ir tenka pastebėti, kad jų labai trūksta. Per paskutinius dešimtmečius kompiuterinei įrangai tampant vis galingesne ir prieinamesne, atsiradus bei ištobulėjus modernioms ryšio technologijoms bei pereinant nuo tradicinės – popierinės – prie elektroninės leidybos, susidarė objektyvios prielaidos pradėti spręsti mokslo žinių valdymo problemą iš naujo.

Vienas labai svarbių mokslo žinių valdymo sudėtinių uždavinių, kurio sprendimui ir skirtas šis straipsnis, yra mokslo informacijos klasifikavimas bendrąja prasme. Šiame darbe apsiribojama kiek siauresne šio uždavinio formuluote – automatizuotas mokslo publikacijų raktinių žodžių radimas. Iš tiesų nesunku suprasti, kad raktinių žodžių, suprantamų kaip keleto svarbiausių straipsnį apibūdinančių mokslo sąvokų, nustatymas yra tam tikra prasme ekvivalentus straipsnio klasifikavimui. Negana to, surastieji raktiniai žodžiai gali būti naudojami ir kitų mokslo informacijos valdymo uždavinių sprendimui. Raktinių žodžių nustatymo automatizavimas išsprendžia nemažai problemų, susijusių su raktinių žodžių kiekybinių bei kokybinių charakteristikų neatitikimais skirtingo laikmečio, leidėjų ir autorių straipsniuose, ir todėl įgalina lankstesnį jų panaudojimą.

*Darbas atliktas pagal tyrimų programą „Mokslo terminų pasiskirstymo specialioje literatūroje stochastinis modeliavimas“, finansuojamą Lietuvos valstybinio mokslo ir studijų fondo (temos nr. G–127).

Straipsnių klasifikavimo algoritmai

Automatizuoto mokslo publikacijų raktinių žodžių radimo problema buvo spręsta ir darbe [1], kuriame apžvelgta keletas paprasčiausių raktinių žodžių ištraukimo (*angl.* keywords extraction) klasei priskiriamų metodų. Šios klasės metodai pasižymi tuo, kad raktiniai žodžiai parenkami iš tekste esančių mokslo terminų, kurie tenkina tam tikras savybes. Atliktas tyrimas parodė, kad būtent dėl šios tiesioginio išrinkimo iš teksto savybės nagrinėtieji metodai nėra efektyvūs mūsų uždaviniui spręsti – didelė dalis (tirtuose duomenyse – maždaug pusė) raktinių žodžių tiesiogiai nesutinkama straipsnio tekste. Todėl sprendimo tenka ieškoti tarp vadinamųjų raktinių žodžių priskyrimo (*angl.* keywords assignment) algoritmų, kurie skiriasi nuo aukščiau aprašytųjų tuo, kad jų kandidatai į raktinius žodžius yra žinomi iš anksto ir nepriklauso nuo konkretaus dokumento teksto.

Tekstų kategorizavimo (klasifikavimo) algoritmai apibrėžia nežinomos funkcijos $f\colon D \times C \to \{0, 1\}$, priskiriančios dokumentus (iš aibės $D = \{d_1, \ldots, d_n\}$) kategorijoms (iš aibės $C = \{c_1, \ldots, c_m\}$), aproksimaciją funkcija $\widehat{f}\colon D \times C \to \{0, 1\}$, kuri vadinama klasifikatoriumi (modeliu, hipoteze). Sutapatinę kategorijas su raktiniais žodžiais, gauname, kad tekstų kategorizavimas ir raktinių žodžių priskyrimas yra ekvivalentūs uždaviniai, būtent todėl raktiniams žodžiams nustatyti dažnai naudojami tekstų kategorizavimo algoritmai. Šiuose algoritmuose dominuoja vadinamasis automatiniu mokymusi (*angl.* machine learning) paremtas klasifikatorių konstravimo būdas – tai atliekama stebint ekspertų atliktų priskyrimų charakteristikas.

Tekstų kategorizavimo algoritmų yra labai daug ([3]), keletas kurių, – LLSF, SVM, kNN, – pasižymi gana geromis efektyvumo charakteristikomis. Visus tris algoritmus su tam tikromis išlygomis galima pritaikyti mūsų uždaviniui spręsti. Taikymas nesukelia jokių sunkumų, tik svarbu atkreipti dėmesį, kad visuose trijuose algoritmuose straipsnius tenka paversti fiksuoto ilgio vektoriais, kuriuose kiekvienas elementas nurodo tam tikro mokslo termino (iš fiksuotos generalinės aibės, kurios dydis lygus vektoriaus ilgiui) svorį straipsnyje. Svoris gali būti tiek binarinis termino buvimo tekste identifikatorius, tiek tam tikra funkcija nuo termino pasikartojimo dažnio, straipsnio ilgio ar kitų charakteristikų.

Nors šie algoritmai deklaruojami kaip efektyvūs ir jie gali būti pritaikyti mūsų uždaviniui spręsti, tačiau yra priežasčių, dėl kurių tenka ieškoti ir kitokių sprendimų. Visų pirma jie kurti kasdienės kalbos tekstams klasifikuoti, todėl juose neišnaudojama mokslinės terminijos specifika. Be to, didžiojoje dalyje tekstų kategorizavimo algoritmų taikomas vadinamasis „juodosios dėžės“ principas, kuris lemia tai, kad nustatyti sąryšiai ir dėsningumai paslepiami nuo naudotojo ir pasireiškia tik per klasifikavimo rezultatus. Tuo tarpu neretai mus domina išreikštinės struktūros, paaiškinančios vieno ar kito sprendimo priėmimą, kas įgalina atlikti ekspertinę analizę ir padaryti pataisymus ar papildymus.

Mokslo terminų aplinkų modeliai

Su mokslo terminu, esančiu pulikacijos tekste, visada galima susieti tam tikrą jį supančią aplinką (kontekstą), kuri neša vienokią ar kitokią informaciją apie šį terminą. Aprašius, kaip iš turimos aplinkos, t.y., konteksto fragmento, išgauti ir panaudoti jos

turimą informaciją apie supamąjį terminą, turėtume įrankį raktinių žodžių nustaymui. M. Hazewinkel siūlo naudoti vadinamuosius identifikacinius debesėlius ([2]), kurie raktinį žodį (kuris yra mokslo terminas) susieja su jo kontekste dažnai sutinkamų mokslo terminų aibe. Ši aibė turėtų būti sudaroma automatiškai (t.y., be dialogo su naudotoju), pakankamai nedidelė (10 – 20 elementų) ir būtinai išreikštinė, tuo sudarant sąlygas srities ekspertams atlikti pataisymus. Sudarytojo debesėlio fragmentas, sutiktas tekste, yra tarsi raktinio žodžio „pėdsakas“, nurodantis, kad kalbama būtent apie šį raktinį žodį (žinoma su tam tikra tikimybe). Toliau pateiksime intuityviai suprantamos identifikacinio debesėlio sąvokos matematinę interpretaciją, paremtą tikimybiniais skirstiniais, bei taikymo raktinių žodžių radimui idėjas.

Tegul T žymi nagrinėjamos mokslo srities mokslo terminų aibę. Kiekvieną srities straipsnį $a \in A$ sutapatiname su jo tekste esančių chronologiškai surikiuotų mokslinių terminų vektoriumi $a = (a_1, \ldots, a_n)$, $a_i \in T$, $n = n(a)$. Manysime, kad straipsnis a susideda iš $q = q(a) \geqslant 1$ nepersikertančių homogeninių dalių, o kiekvieną jų charakterizuoja mokslo terminas ar jų grupė $w_j = w_j(a)$, kurią vadinsime apibendrintu raktiniu žodžiu arba tiesiog raktiniu žodžiu. Pažymėkime:

$$W(a) = \{w_j(a),\ j = 1, \ldots, q(a)\}, \quad W = \bigcup_{a \in A} W(a). \tag{1}$$

Kiekvienam straipsniui $a \in A$ apibrėžta q porų $(N_1, w_1), \ldots, (N_q, w_q)$, kur N_i žymi terminų, sutinkamų homogeninėje straipsnio a dalyje, charakterizuojamoje raktinio žodžio w_i, indeksų aibę. Kartais vietoje porų (N_i, w_i) patogiau naudoti vektorių $(v_1, \ldots, v_n)$, apibrėžiamą lygybe $v_i = w_j$, jei $i \in N_j$.

Kiekvienam raktiniam žodžiui $w \in W$ apibrėšime jo debesėlio sąvoką. Intuityviai tai suvokiama kaip aibė mokslinių terminų, kurių pasirodymo dažnumas w charakterizuojamose straipsnio dalyse ženkliai skiriasi nuo jų pasirodymo dažnumo visoje aibėje A. Vieni šios aibės elementai yra artimesni žodžiui w, kiti tolimesni, todėl tikslinga apibrėžti ir jų svorius. Todėl w debesėlį suprasime kaip tam tikrą funkcionalą $\gamma_w(t), t \in T$, o jo apibrėžimui panaudosime tikimybinių skirstinių terminus.

Tegul N žymi natūraliųjų skaičių aibę, o straipsnis a parenkamas atsitiktinai iš aibės A. Laikydami, kad $v_i = a_i = 0$, kai $i > n$, mes apibrėžiame apibendrintas atsitiktines sekas $\{a_i, i \in N\}$ ir $\{v_i, i \in N\}$ ir galime kalbėti apie daugiamačius pasiskirstymus $P\{a_{i_1} = t_1, \ldots, a_{i_k} = t_k\}$ ir $P\{v_{i_1} = w_1, \ldots, v_{i_k} = w_k\}$.

1 APIBRĖŽIMAS. Tegul τ yra atsitiktinis dydis, priimantis natūraliąsias reikšmes, o a – apibendrintas atsitiktinis dydis, priimantis reikšmes iš aibės A. Santykis

$$\gamma_w(t) = \frac{P\{a_\tau = t | v_\tau = w\}}{P\{a_\tau = t\}} \stackrel{def}{=} \frac{P_w(t)}{P(t)} \tag{2}$$

vadinamas raktinio žodžio $w \in W$ identifikaciniu debesėliu.

Jei funkcionalo γ reikšmės bei stebimo straipsnio a homogeninės dalys $N_1, \ldots, N_q$ yra žinomos, tai priėmus prielaidą apie sekos $\{a_i\}$ sąlyginį stacionarumą bei sąlyginį nepriklausomumą (žr. *prielaidą* 1 žemiau), aibės $W(a)$ vertinimui galima taikyti maksimalaus tikėtinumo metodą (MTM).

1 PRIELAIDA (sąlyginio stacionarumo ir nepriklausomumo). *Tegul visiems natūraliesiems skaičiams* k, $i_1, \ldots, i_k$, *visiems terminams* $t_j \in T$ *ir raktiniams žodžiams* $w \in W$ *galioja lygybė*

$$P\{a_{i_1} = t_1, \ldots, a_{i_k} = t_k | v_{i_1} = \ldots = v_{i_k} = w\} = \prod_{i=1}^{k} P_w(t_i). \tag{3}$$

Tada, galiojant homogeninių dalių $N_1, \ldots, N_q$ *nepriklausomumo prielaidai,* $\widehat{W}(a) = \{\widehat{w}_1, \ldots, \widehat{w}_q\}$, *kur*

$$\widehat{w}_j = \arg\max_{w \in W} \prod_{i \in N_j} P_w(a_i). \tag{4}$$

Aišku, kad formulėje (4) P_w galima keisti į γ_w.

Deja prielaida (3) labai toli nuo tikrovės, todėl geram $W(a)$ įvertinimui reikia naudoti adekvatesnes prielaidas. Visų pirma pateiksime bendresnį identifikacinio debesėlio apibrėžimą.

2 APIBRĖŽIMAS. Tegul $k \in N$, $\tau_1, \ldots, \tau_k$ – nepriklausomi atsitiktiniai dydžiai, priimantys natūraliąsias reikšmes. Santykis

$$\gamma_w(t_1, \ldots, t_k) = \frac{P\{a_{\tau_1} = t_1, \ldots, a_{\tau_k} = t_k | v_{\tau_1} = \ldots = v_{\tau_k} = w\}}{P\{a_{\tau_1} = t_1, \ldots, a_{\tau_k} = t_k\}} \stackrel{def}{=} \frac{P_w(t_1, \ldots, t_k)}{P(t_1, \ldots, t_k)} \tag{5}$$

vadinamas raktinio žodžio $w \in W$ k-tosios eilės identifikaciniu debesėliu.

Pateiktą debesėlio apibrėžimą galima tiesiogiai taikyti nustatant $W(a)$ MTM metodu, tačiau praktikoje labai sudėtinga, o dažnai ir neįmanoma įvertinti $P_w(t_1, \ldots, t_k)$, kai $k \gg 1$. Todėl modelyje siūlomos dvi prielaidos, kurios supaprastina MTM taikymą, tačiau savo ruožtu yra bendresnės už visiško nepriklausomumo prielaidą (3). Paprastumo dėlei prielaidos formuluojamas tuo atveju, kai $N_1, \ldots, N_q$ yra nepersikertantys intervalai.

Kiekvienam natūraliųjų skaičių intervalui $I \subset N$ tegul $H_w(I)$ žymi atsitiktinį įvykį, kad visas I priklauso stebimo straipsnio a homogeninei daliai, charakterizuojamai raktinio žodžio w.

2 PRIELAIDA (sąlyginio markoviškumo). *Visiems intervalams* $I = \{r, r+1, \ldots, m\} \subset N$, *terminams* $t_i \in T$ *ir raktiniams žodžiams* $w \in W$ *galioja lygybė*

$$P\{a_i = t_i, i \in I | H_w(I)\} = P_w(t_r) \cdot \prod_{i=r}^{m-1} P_w(t_{i+1}, t_i) / P_w(t_i). \tag{6}$$

3 PRIELAIDA (homogeninių dalių nepriklausomumo). *Visiems natūraliesiems skaičiams* q *ir* n, *terminams* $t_1, \ldots, t_n \in T$, *raktiniams žodžiams* $w_1, \ldots, w_q \in W$

ir visiems aibės $\{1,\dots,n\}$ *suskaidymams į nepersikertančius intervalus* $N_1,\dots,N_q$ *galioja lygybė*

$$P\left\{a_i = t_i, i = \overline{1,n} \mid H_{w_1}(N_1),\dots,H_{w_q}(N_q)\right\} = \prod_{j=1}^{q} P\left\{a_i = t_i, i \in N_j \mid H_{w_j}(N_j)\right\}. \quad (7)$$

Prielaidomis (6) ir (7) apibrėžtas modelis gali būti taikomos aibės $W(a)$ bei galbūt straipsnio a homogeninių dalių nustatymui MTM metodu ir tam užtenka mokėti įvertinti pirmos bei antros eilės debesėlius:

$$\widehat{w}_1,\dots,\widehat{w}_q,\widehat{N}_1,\dots,\widehat{N}_q\colon \prod_{j=1}^{q}\left[P_{w_j}(t_{r_j})\prod_{i=r_j}^{m_j-1}P_{w_j}(t_{i+1},t_i)/P_{w_j}(t_i)\right]\to\max. \quad (8)$$

Čia laikome, kad $N_j = \{r_j, r_j+1,\dots,m_j\}\subset N$.

Dalis pasiūlytosios identifikacinių debesėlių teorijos buvo išbandyta ir praktiškai – atlikti raktinių žodžių priskyrimo bandymai su realiais duomenimis. Dėl vietos stokos gautų rezultatų nepateiksime, tačiau suformuluosime svarbiausias išvadas: priskyrimo algoritmas, besiremiantis terminų nepriklausomumo ir stacionarumo prielaida (3), yra labai neefektyvus, lyginant su sudėtingesniu algoritmu, nedarančiu prielaidų apie nepriklausomumą bei besiremiančiu apibendrintu debesėlio apibrėžimu (5). Antrasis algoritmas pasižymi viena svarbia savybe – jo efektyvumo priklausomybė nuo debesėlio dydžio yra pakankamai silpna, todėl praktikoje galima naudoti nedidelių dydžių debesėlius. Kita vertus šio algoritmas taikymas reikalauja daug skaičiavimų sąnaudų, todėl artimiausiu metu bus tiriamas raktinių žodžių paieškos metodas, paremtas (8) formule.

Literatūra

1. V. Balys, R. Rudzkis, Mokslinių terminų statistinio pasiskirstymo taikymas straipsnių klasifikavime, *Liet. matem. rink.*, **43**(spec. nr.), 463–467 (2003).
2. M. Hazewinkel, Topologies and metrics of information spaces, *CWI Quaterly*, **12**(2), 93–110 (1999).
3. F. Sebastiani, Machine learning in automated text categorization, *ACM Computing Surveys*, **34**(1), 1–47 (2002).

SUMMARY

V. Balys, R. Rudzkis. Applying models of scientific terms' surroundings in classification of scientific publications

This paper considers a problem of classification of scientific publications wich is treated here as a problem of finding keywords. Keywords are determined by making use of statistical analysis of distribution of scientific terms. Models of scientific terms surroundings – so called identification clouds – are presented as well as results of experiments of their applicability for keywords detection.

Keywords: classification of publications, keywords, statistical models of terms distributions.