Liet. matem. rink, **45**, *spec. nr.*, 2005, 329–334

Stochastinių mokslo terminų aplinkų modelių tyrimas

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

Reziumė. Darbe nagrinėjamas raktinių žodžių priskyrimo mokslo publikacijoms uždavinys. Tiriami ankstesniuose autorių darbuose pasiūlyti mokslo terminų aplinkų stochastiniais modeliais paremti algoritmai, kurie lyginami su populiariomis alternatyviomis procedūromis. Algoritmų efektyvumas vertinamas remiantis realių duomenų pagrindu atliktais tyrimais. Formuluojamos išvados apie siūlomų algoritmų praktines savybes bei taikymų perspektyvas.

Raktiniai žodžiai: identifikaciniai debesėliai, raktiniai žodžiai, klasifikavimo algoritmai.

Įvadas

Darbe sprendžiamas automatinio raktinių žodžių priskyrimo mokslo publikacijoms uždavinys. Uždavinio aktualumą lemia didelės mokslinių žinių, sukauptų bei naujai pateikiamų specialiose publikacijose, apimtys; spartus įprastos popierinės leidybos keitimas elektronine; modernių kompiuterinių priemonių, įgalinančių kaupti, apdoroti bei pateikti naudotojams didelius informacijos archyvus, sukūrimas. Susidomėjimą tokių tyrimų rezultatais reiškia tiek akademinė visuomenė, tiek leidybinį pasaulį atstovaujantys asmenys bei įstaigos.

Šiame darbe pagrindinis dėmesys skiriamas konstruktyvių algoritmų, sukurtų remiantis mokslinio termino aplinkos modeliu, sąvybių tyrimui. Mokslinio termino aplinkos (identifikacinio debesėlio) sąvoką pasiūlė M. Hazewinkel [1]: tai grupė žodžių ar trumpų frazių, dažnai sutinkamų terminą supančiame kontekste. Aplinkos gali būti panaudotos raktiniams žodžiams nustatyti, nes jos yra termino „pėdsakas“ tekste ir signalizuoja apie jo svarbą netgi tuo atveju, kai pats terminas tekste nesutinkamas. Stochastiniai identifikacinių debesėlių modeliai nagrinėjami straipsniuose [3] ir [4].

Pagrindinės sąvokos

Turime fiksuotą mokslo sritį, jos terminų žodyną T bei visų srities straipsnių generalinę aibę A. Kiekvieną straipsnį $a \in A$ sutapatiname su vektoriumi $a = (a_1, \ldots, a_n)$, $a_i \in T$, $n = n(a)$, sudarytu iš chronologine tvarka išvardintų straipsnio tekste esančių mokslo terminų. Straipsnio $a \in A$ raktinių žodžių aibę žymėsime $W(a) = \{w_j(a),\ j = 1 \ldots q(a),\} \subset W$, kur W – fiksuota potencialių mokslo srities raktinių žodžių aibė. *Raktinių žodžių nustatymo uždavinys* – įvertinti nežinomą aibę $W(a)$.

*Darbas atliktas pagal tyrimų programą „Stochastinių mokslo terminų pasiskirstymo specialioje literatūroje modelių tyrimas“, finansuojamą VMSF (temos Nr. G–177).

Raktinių žodžių priskyrimo algoritmo pagrindą sudaro nežinomos klasifikavimo funkcijos $F\colon A \times W \to \{0,1\}$ aproksimavimas įverčiu $\widehat{F}\colon A \times W \to \{0,1\}$. Stochastiniuose modeliuose kiekvienai kategorijai (raktiniam žodžiui) gauto atskiro klasifikatoriaus $\widehat{f}_w\colon A \to R$, $w \in W$ reikšmės interpretuojamos kaip dydis, proporcingas tikimybei, kad w yra atsitiktinio straipsnio a raktinis žodis. $\widehat{W}(a)$ sudaromas iš tų $w \in W$, kuriems $\widehat{f}_w(a)$ yra didžiausi (raktinių žodžių skaičius $q(a)$ turi būti žinomas iš anksto) arba viršija tam tikrą slenkstį (šiuo atveju raktinių žodžių skaičiaus – atsitiktinis dydis, kuris turi būti įvertintas). Taip pat galimi ir kiti būdai.

Tarkime, kad turime mokymo straipsnių imtį A^L, kuriai kiekvienam $a \in A^L$ raktinių žodžių aibė $W(a)$ yra žinoma. Automatinio mokymosi (angl. machine learning) paradigmos algoritmai klasifikatorių f_w aproksimacijas skaičiuoja remdamiesi apmokymo imtyje turimais teisingais klasifikavimo sprendimais. Dažnai skirtingų raktinių žodžių klasifikatoriai yra to pačio pavidalo ir skiriasi tik parametru:

$$\widehat{f}_w(a) = f\big(a, \widehat{\theta}_w(A^L)\big), \quad a \in A, \ \ w \in W, \ \ \theta_w \in \Theta. \tag{1}$$

Čia θ_w – bendru atveju daugiamatis parametras, kurio įvertis skaičiuojamas mokymo imtyje A^L.

Algoritmai

Identifikacinių debesėlių algoritmas (IDA)

Identifikacinių debesėlių (mokslo terminų aplinkų) teorija suformuluota straipsniuose [3] ir [4]. Paprasčiausias siūlomas identifikaciniais debesėliais paremtas algoritmas (IDA) gaunamas iš formulės (4) straipsnyje [3]:

$$\widehat{f}_w(a) = \prod_{a_i \in a} \widehat{\gamma}_w(a_i). \tag{2}$$

Čia $\gamma_w(v)$ žymi termino $v \in T$ pasirodymo atsitiktiniame straipsnyje a tikimybių santykį prie sąlygų $a \in A$ ir $w \in W(a)$ (žr. [3] formulė (2)) o įvertis gaunamas pakeitus tikimybes jų dažnuminiais įverčiais. Jei visus $\widehat{\gamma}_w(a_i)$, išskyrus kelis didžiausius, prilyginsime nuliui, tai gausime norimo dydžio baigtinį debesėlį.

Dažnių algoritmas

Šis algoritmas yra supaprastintas [2] pasiūlyto (8) algoritmo variantas. Jis paremtas intuityvia prielaida, kad jei terminas sutinkamas straipsnio tekste kur kas dažniau, nei įprastai, tai jis greičiausiai yra to straipsnio raktinis žodis:

$$\widehat{f}_w(a) = \frac{d(w,a)}{d(w)}, \quad d(w) = \frac{1}{|A^L|} \sum_{\substack{b \in A^L \\ d(w,b)>0}} d(w,b). \tag{3}$$

Čia $d(w,a)$ – pasirinktu būdu reprezentuojamas termino w svoris straipsnio a tekste, o $|B|$ žymi aibės B elementų skaičių. Algoritmas veikia tik tiems raktiniams žodžiams, kurie tiesiogiai sutinkami straipsnių tekstuose.

K artimiausių kaimynų metodas (kNN)

Tai vienas žinomiausių ir daugelyje sričių taikomas algoritmas:

$$\widehat{f}_w(a) = \frac{1}{k} \sum_{b \in K(a)} f_w(b). \tag{4}$$

Čia $K(a) = K(a,k)$ žymi k artimiausių straipsnio a kaimynų mokymo imtyje A^L aibę, $f_w(b)$ – žinomą klasifikatoriaus reikšmę atitinkamam tos aibės elementui. Artimiausių kaimynų aibė priklauso nuo dviejų straipsnių atstumo mato parinkimo, pvz.:

$$\rho(a_1, a_2) = \sum_{t \in T} \left| d(t, a_1) - d(t, a_2) \right|^2, \quad a_1, a_2 \in A. \tag{5}$$

Tiesinis mažiausių kvadratų metodas (LLSF)

LLSF metodas (ang. Linear Least Squares Fit), remiasi prielaida, kad raktinių žodžių svoriai yra tiesinė straipsnio terminų svorių kombinacija. Jei kiekvienas straipsnis $a \in A$ užrašomas vienodo ilgio vektoriais $D(a) = (d_1(a), \ldots d_{|T|}(a))$, kur $d_i(a)$ yra i-ojo termino iš sunumeruotos aibės T svoris straipsnyje a, o klasifikatorių reikšmės pateikiamos analogišku vektoriumi $\widehat{C}(a) = (\widehat{f}_1(a), \ldots \widehat{f}_{|W|}(a))$ (sunumeravus aibės W elementus), tai atsitiktinio straipsnio $a \in A$ klasifikavimas aprašomas lygybe:

$$D(a)\widehat{B} = \widehat{C}(a), \tag{6}$$

kur nežinomos matricos B įvertis $\widehat{B}$ yra gaunamas iš sąlygos:

$$D(A^L)\widehat{B} = C(A^L). \tag{7}$$

Čia matricos $D(A^L)$ ir $C(A^L)$ yra žinomos – jos sudarytos iš apmokymo imties straipsnius atitinkančių vektorių.

(7) sprendimo nedetalizuosime dėl vietos stokos, tik paminėsime, kad jis reikalauja labai daug skaičiavimų. Nesunku įžvelgti analogiją tarp matricos $\widehat{B}$ stulpeliuose esančių terminų „indėlių“ į raktinio žodžio svorį straipsnyje bei identifikacinių debesėlių, todėl ir čia naudosime identifikacinio debesėlio sąvoką. Siekdami dar padidinti panašumą, įvesime fiksuoto dydžio debesėlio reikalavimą, t.y. beveik visi kiekvieno matricos $\widehat{B}$ stulpelio elementai, išskyrus kelis (algoritmo parametras) didžiausius, prilyginami nuliui.

Eksperimentinė dalis

Ankstesniame skyrelyje išvardintus raktinių žodžių priskyrimo algoritmus tyrėme atlikdami bandymus su 2132 kompiuterijos srities straipsnių santraukomis. Kadangi turėjome tik santraukas, identifikacinių debesėlių algoritme apsiribota vienos straipsnio homogeninės dalies atveju. Dėl duomenų stokos tik 19 raktinių žodžių klasifikatorių buvo įmanoma apmokyti, o kiekviename iš straipsnių (išskyrus labai nedidelę dalį) liko tik po vieną raktinį žodį. Tai savo ruožtu lėmė dviejų paprastų klasifikavimo strategijų parinkimą: a) straipsniui priskiriamas lygiai vienas raktinis

žodis bei b) straipsniui priskiriami lygiai du raktiniai žodžiai. Abiem atvejais priskyrimas atliekamas pagal didžiausias klasifikavimo funkcijų reikšmes. Visos aukščiau išvardintos priežastys sąlygoja tai, kad rezultatai turėtų būti interpretuojami kaip pirminis siūlomų algoritmų patikrinimas, siekiant išsiaiškinti jų tolimesnio tobulinimo ir naudojimo perspektyvą, o ne kaip galutiniai ir patikimi rezultatai, pagal kuriuos būtų galima daryti griežtas išvadas.

Algoritmų efektyvumui įvertinti bei palyginti apmokymo imtį A^L padalinome į dvi nesikertančias dalis, kurių pirmojoje konstruojamas klasifikatorius, t.y. vertinamas parametras θ, o antrojoje vertinamas efektyvumas. Algoritmų efektyvumui vertinti naudotas matas Re (angl. recall – žr. [5]), kuris parodo, kurią dalį autoriaus nurodytų raktinių žodžių surado algoritmas.

Dažnių ir kNN algoritmuose svorių funkcija $d(w,a)$ buvo lygi populiariajam termino ir dokumento ryšio matui $tfidf(w,a)$ ([5]). Pirmoje lentelėje pateikti algoritmų efektyvumo rezultatai atspindi optimalų algoritmų efektyvumą, kuris gaunamas bandymų keliu parinkus tinkamas algoritmų parametrų reikšmes. Mes naudojome tokias parametrų reikšmes: 25 artimiausi kaimynai kNN algoritme, 50 elementų debesėliai LLSF procedūroje bei 100 elementų debesėliai IDA algoritme.

Tyrimų rezultatai ir išvados

1 lentelėje pateiktas algoritmų efektyvumo palyginimas.

2 ir 3 stulpeliuose pateikti efektyvumo mato Re empiriniai įverčiai atitinkamai pirmai klasifikavimo strategijai (vienas raktinis žodis) ir antrai (du raktiniai žodžiai). Matome, kad geriausius rezultatus duoda IDA bei LLSF algoritmai ir natūraliai efektyvesnis yra antrasis, paremtas kur kas sudėtingesniais skaičiavimais. Artimiausių kaimynų algoritmo efektyvumas netikėtai labai mažas, tačiau tai greičiausiai galima pagrįsti apmokymo duomenų specifika: kaimynai randami pagal pakankamai trumpus teksto fragmentus (santraukas). Dažnių skaičiavimu paremtas algoritmas duoda neprastus rezultatus, žinant, kad jis raktinius žodžius parenka tik iš tiesiogiai tekste sutinkamų terminų. Jeigu likusiusiems algoritmams uždėtume tokį pat apribojimą, t.y. straipsnio raktinius žodžius jie galėtų parinkti tik iš tų pretendentų, kurie sutinkami tekste, gautume 4 ir 5 stulpelyje pateiktus atitinkamų strategijų įverčius $\widehat{Re}'$, iš kurių matome, kad tik LLSF algoritmas išlieka toks pat efektyvu s, kaip dažnių algoritmas. Matome, kad didžiausią efektyvumo prieaugį (atsisakius apribojimo) duoda IDA bei LLSF algoritmai. Šis efektyvumo prieaugis yra vienas iš identifikaciniais debesėliais paremtų algoritmų naudojimo pagrindimų.

1 lentelė

Algoritmas / Matas	$\widehat{Re}_1$	$\widehat{Re}_2$	$\widehat{Re}'_1$	$\widehat{Re}'_2$
kNN	0.33	0.48	0.25	0.30
Dažnių	0.48	0.50	**0.48**	**0.50**
IDA	0.53	0.66	0.32	0.39
LLSF	**0.67**	**0.83**	**0.45**	**0.51**

2 lentelė

Debesėlio dydis / Įvertis	$\widehat{Re}_1^{IDA}$	$\widehat{Re}_2^{IDA}$	$\widehat{Re}_1^{LLSF}$	$\widehat{Re}_2^{LLSF}$
5	0.27	0.42	0.67	0.81
10	0.31	0.46	0.67	0.81
20	0.39	0.52	0.67	0.81
50	0.48	0.61	0.68	0.83

2 lentelėje pateikiami IDA bei LLSF algoritmų efektyvumo įverčio $\widehat{Re}$ priklausomybės nuo debesėlių dydžio tyrimo rezultatai, kurie gauti pakartotinai taikant šiuos algoritmus tiems patiems duomenims, bet su skirtingomis debesėlių dydžių reikšmėmis:

Akivaizdus LLSF algoritmo pranašumas prieš naivia terminų nepriklausomumo prielaida paremtą IDA algoritmą: jo efektyvumas nepriklauso nuo debesėlio elementų skaičiaus, ir netgi su 5 ar 10 elementų debesėliais galima sėkmingai atlikti raktinių žodžių priskyrimą.

Visgi LLSF algoritmas turi ir trūkumų: jo debesėlių sudarymui reikalingi daug kompiuterio resursų reikalaujantys skaičiavimai, be to jis neleidžia panaudoti informacijos apie straipsnio tekste sutinkamų terminų tarpusavio padėtį. Norėtusi turėti gal ir ne tokius efektyvius, tačiau reikalaujančius mažiau skaičiavimų algoritmus. Būtina ištirti sudėtingesnius identifikaciniais debesėliais pagrįstus algoritmus ([3], [4]), kuriuose taikomos silpnesnės nei visiško terminų nepriklausomumo prielaidos, iteratyvios straipsnio homogeninių dalių nustatymo procedūros, modelių parametrizacija ir kiti aspektai, nes intuityviai suprantamą terminų nepriklausomumo prielaidos neadekvatumą patvirtino ir šie atlikti tyrimai. Tyrimus atlikti bei jų rezultatus publikuoti numatoma artimiausiu metu, gavus tinkamus duomenis.

Literatūra

1. M. Hazewinkel, Topologies and metrics of information spaces, *CWI Quaterly*, **12**(2), 93–110 (1999).
2. V. Balys, R. Rudzkis, Mokslinių terminų statistinio pasiskirstymo taikymas straipsnių klasifikavime, *Liet. matem. rink.*, **43** (spec. nr.), 463–467 (2003).
3. V. Balys, R. Rudzkis, Mokslo terminų aplinkų modelių taikymas straipsnių klasifikavime, *Liet. matem. rink.*, **44** (spec. nr.), 537–541 (2004).
4. V. Balys, R. Rudzkis, Stochastic models for keyphrase assignment, in: *Proceedings of the Seventh International Conference Computer Data Analysis and Modeling*, vol. 2 (2004), pp. 118–122.
5. F. Sebastiani, Machine learning in automated text categorization, *ACM Computing Surveys*, **34**(1), 1–47 (2002).
6. Y. Yang, C.G. Chute, A linear least squares fit mapping method for information retrieval from natural language texts, in: *Proceedings of COLING-92, the 15th International Conference on Computational Linguistics* (1992).

SUMMARY

V. Balys, R. Rudzkis. Analysis of stochastic models of identification clouds

This paper deals with problem of automatic keywords assignment for scientific publications. In recent papers proposed identification clouds based classification algorithms are analysed and compared to popular alternavite methods. The efficiency of algorithms is estimated through real data based experiments. Conclusions about practical aspects and applicability of proposed algorithms are drawn.

Keywords: identification clouds, keywords, classification algorithms.