Lietuvos matematikos rinkinys
Lietuvos matematikų draugijos darbai, ser. B
56 t., 2015, 1–6

ISSN 0132-2818
www.mii.lt/LMR/

taikant latentinį Dirichlė paskirstymo modelį

Darius Aliulis, Vytautas Janilionis

Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas
Studentų 50, LT-51386 Kaunas
E. paštas: darius.aliulis@ktu.edu, vytautas.janilionis@ktu.lt

Santrauka. Vartotojų segmentavimas yra vienas iš aktualiausių reklamos internete uždavinių. Darbe šio uždavinio sprendimui siūloma taikyti adaptyvų latentinį Dirichlė paskirstymo modelį, kuris tinka ir didžiųjų duomenų rinkinių analizei. Pasiūlyti reklamų priskyrimo segmentams kriterijai ir ištirtos segmentavimo kokybės metrikų priklausomybės nuo segmentų skaičiaus.

Raktiniai žodžiai: latentinis Dirichlė paskirstymas, Bajeso metodas, didžiųjų duomenų rinkinių analizė, vartotojų segmentavimas, reklama internete.

Įvadas

Reklama internete yra labai sparčiai auganti paslaugų sritis. Lyginant su kitomis reklamos rūšimis, ji leidžia efektyviausiai pasiekti vartotojus. Vienas iš aktualiausių jos uždavinių yra tikslinės reklamos pateikimas konkrečiam vartotojų segmentui. Vartotojų elgsena internete grįstoje reklamoje (angl. *behavioral targeting*), vartotojų segmentavimo uždavinio sprendimui, dažniausiai taikomi klasikiniai klasterizavimo metodai [9]. Paskutiniais metais paskelbti keli darbai, kuriuose siūloma naudoti teksto turinio analizės modelius [8, 7, 4], tačiau tyrimai atlikti tik su mažu segmentų kiekiu, mažomis imtimis ir neatsižvelgta į mažų segmentų, pvz., turinčių tik vieną vartotoją, įtaką rezultatams.

Šio tyrimo objektas – reklamos internete vartotojų segmentavimo modeliai ir reklamų priskyrimo segmentams kriterijai. Tyrimo tikslas – pritaikyti adaptyvų latentinį Dirichlė paskirstymo modelį reklamos internete vartotojų segmentavimui, pasiūlyti reklamų priskyrimo vartotojų segmentams kriterijus ir ištirti segmentavimo kokybės metrikų priklausomybes nuo segmentų skaičiaus.

1 Latentinis Dirichlė paskirstymo modelis

Tarkime, turime D reklamos internete vartotojų. Juos reikia suskirstyti į K segmentų pagal jų interneto paieškos užklausas, kurios sudarytos iš V ilgio žodyno žodžių. Vartotojų užklausų žodžiai aprašomi generuojančiu latentiniu Dirichlė paskirstymo modeliu [2]:

1. Kiekvienam vartotojų segmentui $k \in \{1, \ldots, K\}$ sugeneruojamas V ilgo žodžių tikimybių vektorius $\beta_k \sim \mathrm{Dir}(\eta)$, čia η simetrinio Dirichlė skirstinio apriorinis parametras.

2. Kiekvienam vartotojui $d \in \{1, \ldots, D\}$, kuris apibrėžiamas jo pateiktų interneto paieškos užklausų žodžių vektoriumi $\mathbf{w}_d = \{w_{d1}, \ldots, w_{dn}\}$:
   - sugeneruojamas K ilgio vartotojo priskyrimo segmentams tikimybių vektorius $\theta_d \sim \mathrm{Dir}(\alpha)$.
   - Kiekvienam vartotojo d pateiktų užklausų žodžiui $w_{dn}, n \in \{1, \ldots, N\}$:
     – sugeneruojamas žodžio segmento numeris, pasiskirstęs pagal polinominį skirstinį $z_{dn} \sim \mathrm{Multi}(\theta_d)$, $z_{dn} \in \{1, \ldots, K\}$;
     – sugeneruojamas užklausos žodis $w_{dn} \sim \mathrm{Multi}(\beta_{z_{dn}})$, $w_{dn} \in \{1, \ldots, V\}$.

Latentinio Dirichlė paskirstymo modelio pilnoji tikimybė [2]:

$$p(\theta, \beta, \mathbf{z}, \mathbf{w}|\alpha, \eta) = \prod_{k=1}^{K} p(\beta_k|\eta) \left(\prod_{d=1}^{D} p(\theta_d|\alpha) \left(\prod_{n=1}^{N} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta_{z_{dn}}) \right) \right).$$

Nežinomų aposteriorinio skirstinio parametrų θ, β, z įvertinimui taikomas variacinis Bajeso metodas [2, 5]. Aposteriorinio skirstinio kintamieji yra susieti, tačiau variacinio skirstinio (1) kintamieji yra nepriklausomi ir kiekvienas yra susietas su skirtingu variaciniu parametru (λ, γ, ϕ) [2, 5]:

$$q(\theta, \mathbf{z}, \beta) = \prod_{k=1}^{K} q(\beta_k|\lambda_k) \prod_{d=1}^{D} \left(q(\theta_d|\gamma_d) \prod_{n=1}^{N} q(z_{dn}|\phi_{dn}) \right). \tag{1}$$

Latentinio Dirichlė paskirstymo modelio tikėtinumo funkcija ir jos skaičiavimo algoritmas pateiktas [2]. Variacinio Bajeso metodo iteracija:

1. Kiekvienam segmentui $k \in \{1, \ldots, K\}$ ir žodyno žodžiui $v \in \{1, \ldots, V\}$, skaičiuojama $\lambda_{kv}^{(t+1)} = \eta + \sum_{d=1}^{D} \sum_{n=1}^{N} 1(w_{dn} = v) \phi_{dnk}^{(t)}$.
2. Kiekvienam vartotojui $d \in \{1, \ldots, D\}$ skaičiuojama $\gamma_{dk}^{(t+1)} = \alpha_k + \sum_{n=1}^{N} \phi_{dnk}^{(t)}$.
   - Kiekvienam vartotojo pateiktų paieškos užklausų žodžiui $n \in \{1, \ldots, N\}$ skaičiuojama $\phi_{dnk}^{(t+1)} \propto \exp\left\{ \Psi\left(\gamma_{dk}^{(t+1)}\right) + \Psi\left(\lambda_{kw_{dn}}^{(t+1)}\right) - \Psi\left(\sum_{v=1}^{V} \lambda_{kv}^{(t+1)} \right) \right\}$, čia Ψ yra digama funkcija (pirmoji funkcijos $\log \Gamma$ išvestinė).

Šiame darbe interneto vartotojų segmentavimui pasiūlyta taikyti adaptyvų variacinį Bajeso metodą [5], kuris leidžia modelį taikyti didžiųjų duomenų rinkinių analizei. Variacinis parametras λ apskaičiuojamas pagal

$$\lambda^{(t+1)} = \left(1 - t^{-0,3}\right) \lambda^{(t)} + t^{-0,3} \widetilde{\lambda}^{(t)}. \tag{2}$$

Taikant stochastinius optimizavimo metodus parametrų vertinimui dažnai naudojami mini-paketai

$$\widetilde{\lambda}_{kv}^{(t+1)} = \eta + \frac{D}{S} \sum_{s=1}^{S} \sum_{n=1}^{N} 1(w_{sn} = v) \phi_{snk}^{(t)}, \quad S > 1, \tag{3}$$

čia S yra vartotojų skaičius t mini-pakete.

2 Segmentavimo kokybės metrikos

Toliau naudojami šie žymėjimai: $A = \{a_1, a_2, \ldots, a_I\}$ yra reklamų aibė; $U_i = \{u_{i1}, u_{i2}, \ldots, u_{im_i}\}$ yra aibė vartotojų, kuriems parodyta reklama a_i; $\varphi(u_{ij})$ yra a_i reklamos parodymų skaičius j-ajam vartotojui; $\delta(u_{ij})$ yra reklamos a_i paspaudimų skaičius, kuriuos atliko j-asis vartotojas; m_i yra skaičius vartotojų, kuriems buvo parodyta reklama a_i.

Segmentavimo kokybės vertinimui naudojamos šios metrikos:

Tikslumas (P) apibūdina į kokią dalį reklamos a_i parodymų g_k segmento vartotojai sureaguos paspaudimais.

$$P(a_i|g_k) = \frac{\sum_{u_{ij} \in g_k(U_i)} \delta(u_{ij})}{\sum_{u_{ij} \in g_k(U_i)} \varphi(u_{ij})}. \tag{4}$$

čia $g_k(U_i)$ yra reklamos a_i segmento vartotojai.

Atkuriamumas (R) apibūdina kokią dalį reklamos a_i paspaudimų atlieka g_k segmentui priskirti vartotojai:

$$R(a_i|g_k) = \frac{\sum_{u_{ij} \in g_k(U_i)} \delta(u_{ij})}{\sum_{j=1}^{m_i} \delta(u_{ij})}. \tag{5}$$

F-matas (F) yra tikslumo ir atkuriamumo harmoninis vidurkis:

$$F(a_i|g_k) = \frac{2 \cdot P(a_i|g_k) \cdot R(a_i|g_k)}{P(a_i|g_k) + R(a_i|g_k)}. \tag{6}$$

Darbe pasiūlyta reklamos a_i priskyrimui segmentams naudoti F-matą arba segmentų aibę apriboti pagal vidutinį atkuriamumą

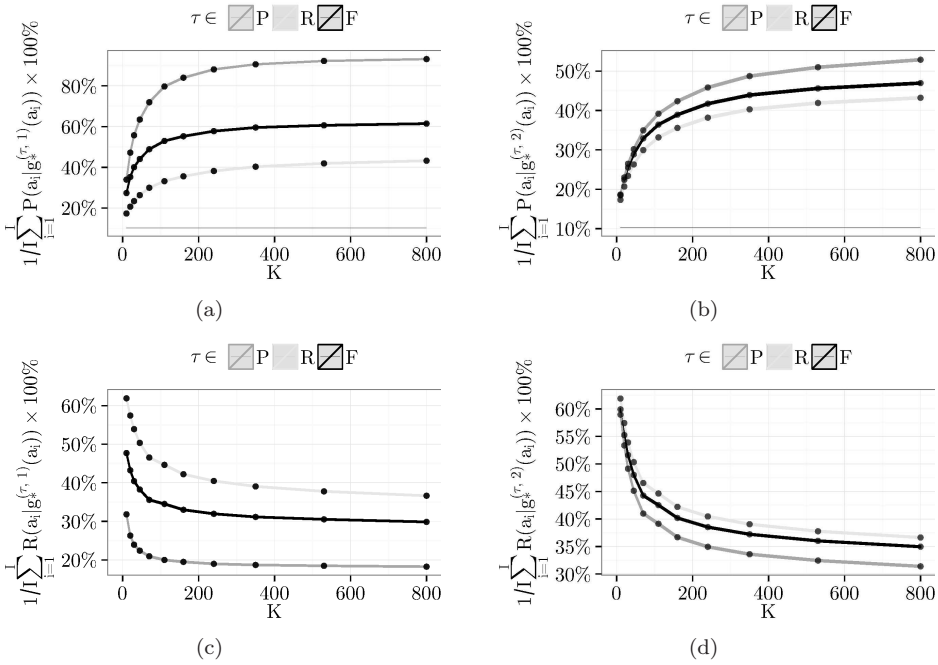
$$g_*^{(\tau,\nu)}(a_i) = \arg\max_{\widetilde{g}_k(\nu)} \tau\big(a_i|\widetilde{g}_k(\nu)\big), \tag{7}$$

čia $k \in \{1, 2, \ldots, K\}$, $\tau \in \{P, R, F\}$, $\nu \in \{1, 2\}$, $\widetilde{g_k}(1) \in \{g_k\}$ yra visų reklamos a_i segmentų aibė, $\widetilde{g}_k(2) \in \{g_k|\ R(g_k) \geq \overline{R(G(U_i))}\}$ yra segmentų aibės apribojimas pagal vidutinį atkuriamumą, $\overline{R(G(U_i))}$ yra reklamos a_i segmentų atkuriamumo vidurkis.

3 Tyrimo rezultatai

Panaudojus didžiųjų duomenų rinkinių analizės programines priemones *BIDMach* [3] ir *Apache Spark* [1] sukurta programa, kuria atlikti tyrimai su realiais reklamos internete duomenimis [6]. Po duomenų filtravimo (pašalinti pateikę labai mažai užklausų arba nepaspaudę nei vienos reklamos vartotojai, neindentifikuoti vartotojai ir t. t.) duomenų rinkinyje liko 66 462 376 reklamų parodymų ir paspaudimų įrašai, $D = 5\,647\,787$ skirtingi vartotojai, $I = 105\,745$ skirtingos reklamos.

Matome (1(c) pav.), kad reklamas priskiriant segmentams pagal tikslumą (4) ir netaikant apribojimų, kai $K = 160$, vidutinė atkuriamumo (5) reikšmė yra mažesnė už 19,6%, kai $K = 800$ – mažesnė už 18,3%, o priskiriant pagal F-matą (6), kai $K = 800$, vidutinė atkuriamumo reikšmė yra 28,6%, o vidutinė tikslumo reikšmė yra 61,4% (1(a) pav.). Kai taikomas apribojimas pagal vidutinį atkuriamumą ir



1 pav. Reklamos internete vartotojų segmentavimo kokybės metrikų priklausomybės nuo segmentų skaičiaus.

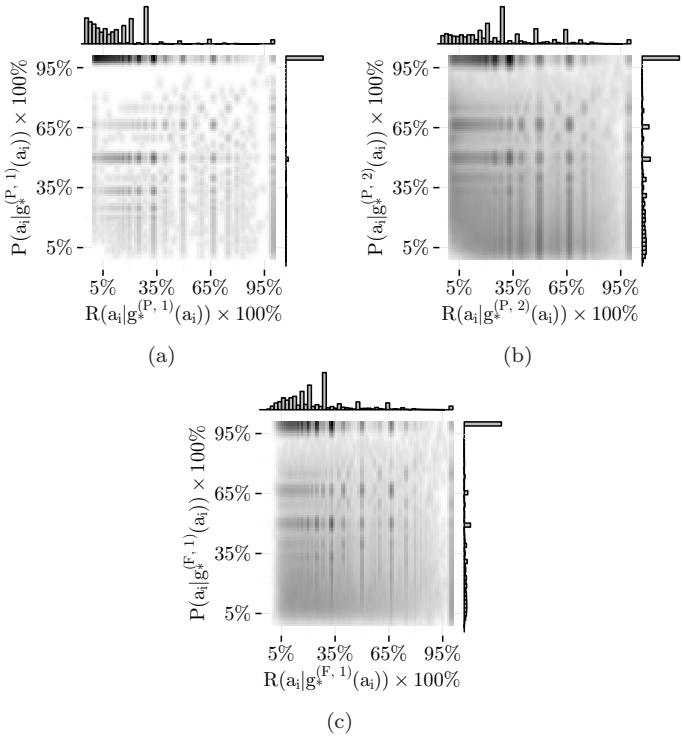
naudojamas tikslumo kriterijus (1(d) pav.), kai $K = 800$, vidutinė atkuriamumo reikšmė yra 31,4%, o vidutinė tikslumo reikšmė yra 52,9% (1(b) pav.).

2 pav. pateikti $I = 105\,745$ reklamų tikslumo (4) ir atkuriamumo (5) metrikų dvimačio pasiskirstymo tankio branduoliniai įverčiai ir marginalios histogramos, kai branduolio funkcija yra Gauso, $K = 800$. Reklamos segmentams priskiriamos pagal: (a) tikslumą, kai netaikomi apribojimai, (b) tikslumą, kai taikomas apribojimas pagal vidutinį atkuriamumą, (c) pagal F-matą, kai netaikomi apribojimai.

4 Išvados

Sudarytas adaptyvus latentinis Dirichlė paskirstymo modelis leidžia geriau segmentuoti interneto vartotojus naudojant jų paieškos užklausų duomenis. Pasiūlyti reklamų priskyrimo segmentams kriterijai, įvertinantys mažų segmentų įtaką rezultatams.

Atliktas tyrimas su realiais interneto vartotojų paieškos užklausų, reklamų parodymų ir paspaudimų duomenimis parodė, kad tikslumo kriterijus netinka reklamų priskyrimui segmentams, kai netaikomi apribojimai, nes tuomet gaunama maža vidutinė atkuriamo reikšmė. Abibendrinant rezultatus galima teikti, kad sprendžiant reklamos internete vartotojų segmentavimo uždavinį ir siekiant geresnės segmentavimo kokybės, reikia taikyti F-mato kriterijų, kai netaikomas apribojimas pagal vidutinį atkuriamumą arba tikslumo kriterijų, kai taikomas apribojimas pagal vidutinį atkuriamumą.



2 pav. Tikslumo ir atkuriamumo metrikų dvimačio pasiskirstymo tankio branduoliniai įverčiai, kai $K = 800$.

Literatūra

[1] M. Armbrust, R.S. Xin, C. Lian, Y. Huai, D. Liu, J.K. Bradley, X. Meng, T. Kaftan, M.J. Franklin, A. Ghodsi and M. Zaharia. Spark SQL: Relational data processing in Spark. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pp. 1383–1394, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-2758-9.

[2] D.M. Blei, A.Y. Ng and M.I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, **3**:993–1022, March 2003. ISSN 1532-4435.

[3] J. Canny and H. Zhao. Big data analytics with small footprint: squaring the cloud. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 95–103. ACM, 2013.

[4] X. Gong, X. Guo, R. Zhang, X. He and A. Zhou. Search behavior based latent semantic user segmentation for advertising targeting. In *2013 IEEE 13th International Conference on Data Mining (ICDM)*, pp. 211–220, Dec 2013.

[5] M. Hoffman, D.M. Blei and F. Bach. Online learning for latent dirichlet allocation. *Adv. Neur. Inf. Proc. Syst.*, **23**:856–864, 2010.

[6] KDD Cup. KDD Cup 2012 Track2 Dataset (interaktyvus) [žiūrėta 2015-06-25]. Adresas internete: http://www.kddcup2012.org/c/kddcup2012-track2.

[7] S. Tu and C. Lu. Topic-based user segmentation for online advertising with latent dirichlet allocation. In L. Cao, J. Zhong and Y. Feng(Eds.), *Advanced Data Mining and*

Applications, volume 6441 of *Lect. Not. Comp. Sci.*, pp. 259–269. Springer, Berlin, Heidelberg, 2010. ISBN 978-3-642-17312-7.

[8] X. Wu, J. Yan, N. Liu, S. Yan, Y. Chen and Z. Chen. Probabilistic latent semantic user segmentation for behavioral targeted advertising. In *Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising*, ADKDD '09, pp. 10–17, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-671-7.

[9] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang and Z. Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pp. 261–270, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-487-4.

SUMMARY

User segmentation in online advertising using latent dirichlet allocation
D. Aliulis, V. Janilionis

User segmentation is one the most important problems in online advertiting. The use of online latent Dirichlet allocation model for analysing big datasets for this purpose is proposed in this paper. The relationship between the number of segments and segmentation quality metrics is analized and criteria for assigning ads to user segments are proposed.

Keywords: latent Dirichlet allocation, bayesian method, big data analytics, user segmentation, online advertising.