Hierarchické prístupy k modelovaniu témy v dokumentoch

Miroslav Smatana, Peter Butka, Matúš Gore

Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach
Letná 9, 042 00 Košice, Slovenská republika

{miroslav.smatana, peter.butka}@tuke.sk,
matus.gore@student.tuke.sk

Abstrakt. Digitálne textové dáta predstavujú v dnešnej dobe dôležitý zdroj informácií. Avšak v súčasnosti je ich počet taký obrovský, že ich manuálne spracovanie a extrakcii informácií by bola časovo veľmi náročná. Existuje niekoľko spôsobov automatickej analýzy textových dát, jednou z nich je modelovanie témy, ktoré ponúka nové možnosti na vyhľadávanie, prehľadávanie a sumarizáciu textových dokumentov. Preto je hlavným cieľom tohto článku predstaviť rôzne metódy hierarchického modelovania téma a taktiež porovnanie vybraných modelov z pohľadu kvality budovania hierarchie tém.

Kľúčové slová: modelovanie témy, hierarchia, Latentná Dirichletova Alokácia

1 Úvod

V súčasnosti s príchodom sociálnych sieti, online časopisov a novín, je veľké množstvo textových dát v digitálnej podobe. Tie predstavujú zaujímavý a dôležitý zdroj informácií. Napríklad na sociálnej sieti Twitter je denne publikovaných okolo 340 miliónov príspevkov, ktoré zvyčajne odrážajú používateľov postoj na niektorú zo svetových udalostí, produkt, či osoby.

Takéto dáta nachádzajú svoje využitie najmä v marketingu, pretože marketing bol stále závislý na dátach a ich správne pochopenie a použitie môže firme priniesť konkurenčnú výhodu a zlepšiť jej postavenie na trhu. Využitie digitálnych textových dát (najmä príspevky zo sociálnych sietí) je možné použiť napríklad pri:

- analýze krízových situácií - príkladom je obdobie vojny, kedy je z týchto dát možné získať reakcie ľudí na danú situáciu;
- zavedenie nového produktu na trh - monitorovanie reakcií používateľov, čo sa im na produkte páči, aké ma chyby;
- v médiách - je možné sledovať o čom ľudia na internete najčastejšie píšu a čo ich zaujíma.

Ako je možné vidieť, analýzu digitálnych textových dát nemožno podceňovať. Avšak ako už bolo spomenuté existuje ich veľké množstvo a preto ich manuálne spracova-

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 136-140.

nie, analýza a extrakcia vhodných informácií by bola veľmi časovo náročná. Preto je potrebné tento proces automatizovať.

V súčasnosti existuje niekoľko druhov analýz, ktoré tento problém riešia. Jednou z nich je modelovanie tém, ktoré umožňuje automatickú analýzu textových príspevkov a predstavuje nový spôsob ich hľadania, prehľadávania a sumarizácie. Hlavným cieľom modelovania tém je vytváranie skupín slov (tém), ktoré sa často vyskytujú spoločne vo vstupnej množine textových dokumentov.

Preto sa v článku sa zameriame na prehľad metód modelovania témy a porovnanie metód hierarchického modelovania témy, ktoré ponúka podrobnejšiu analýzu ako klasické metódy.

2 Modelovanie témy

V tejto sekcii predstavíme modelovanie témy. Ako už bol povedané, cieľom modelovania tém je vytváranie skupín (tém), ktoré sa spolu vyskytujú dostatočne často, pomocou hľadania skrytých tematických štruktúr vo vstupnej kolekcii dokumentov. Ako jednou z prvých metód, ktorá sa pokúšala riešiť tento problém môže byť chápaná latentná sémantická analýza [1], ktorá sa snaží odhaľovať skryté sémantické štruktúry z textov. Formálne však nie je predstavovaná ako jedna z metód modelovania témy, ale bola základom pre ďalšie rozšírenia ako pravdepodobnostná latentná sémantická analýza [2], ktorá tvorí základ pre Latentnú Dirichletovu Alokáciu (LDA) [3]. LDA predstavuje jednu z najpoužívanejších metód v oblasti modelovania tém. Z toho dôvodu sa stala základom pre ďalšie metódy modelovania tém [4,5]. Okrem toho vzniklo niekoľko metód, nezávislých od LDA [6,7,8].

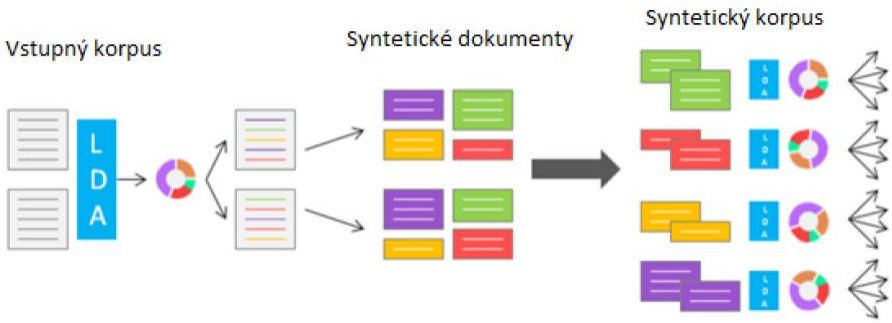
Spomínané práce sa zameriavali na spracovanie dlhých dokumentov. Treba však povedať, že na dátach ako sú príspevky zo sociálnych sieti, kde tieto správy obsahujú len veľmi krátke a stručné texty, nedosahujú dobré výsledky. Preto bolo vyvinutých niekoľko metód [9,10,11], ktoré sú schopne pracovať s krátkymi textami.

Ako je možné vidieť, modelovanie témy je dobre preštudovaná oblasť výskumu, avšak v súčasnosti už tieto štandardné metódy nemusia používateľovi poskytovať dostatočné množstvo informácií. Preto vzniklo niekoľko metód zameraných na budovanie hierarchie tém, ktoré poskytuje podrobnejšie informácie. Jednou z prvých takýchto metód bolo hierarchické LDA (hLDA) [13] a práca [12] v ktorej autori na budovanie tém využívajú proces vnorenej čínskej reštaurácie. Ďalšími z metód hierarchického modelovania témy sú Pachinov alokačný model [14] a metóda HLTA [15].

2.1 hLDA

Metóda hLDA predstavuje jednu z najjednoduchších metód budovania hierarchie tém. Táto metóda využíva prístup "zhoda-nadol" pre vytváranie hierarchie tém pomocou rekurzívneho rozdeľovania a znovu modelovania korpusu pomocou klasické LDA. Výsledkom LDA je model, kde každé slovo v dokumente je priradené téme. HLDA využíva túto informáciu na vytváranie nových syntetických dokumentov pre každú

nájdenú tému zo vstupných dokumentov. Syntetické dokumenty obsahujú iba slová priradené danej téme. Celý tento proces je znázornený na **Obr. 1**.



Obr. 1. Proces vytvárania tém pomocou hLDA

3 Porovnanie metód hierarchického modelovania tém

V tejto kapitole popíšeme porovnanie nami implementovanej metódy hLDA s metódou HLTA. Zvolené metódy sme porovnávali na datasete 20 Newsgroups[1], ktorý obsahuje 20000 dokumentov rozdelených do 20 rôznych tém.

Kvalitu metód sme porovnávali na základe rýchlosti budovania hierarchie a taktiež na základe "coherence score" [16]. V **Tab. 1** je možné vidieť rýchlosť vytvárania hierarchie pre metódy hLDA a HLTA na rôznych vzorkách zo vstupného datasetu (na každej zo vzoriek bolo vykonané jedno meranie). Ako je možné vidieť metóda hLDA je omnoho pomalšia a s narastajúcim počtom vstupných dokumentov sa čas jej spracovania drasticky zvyšuje. Z pohľadu "coherence score" nám, na vzorke 30 dokumentov z 20 Newsgroups datasetu, pre hLDA vyšla hodnota -14.6412 a pre HLTA hodnota 12.6328. Na základe týchto hodnôt je možne povedať, že na zvolenom datasete dosahovali tieto metódy porovnateľné výsledky z hľadiska kvality vytvorených tém.

Tab. 1. Porovnanie metód hierarchického modelovania tém na základe rýchlosti budovania hierarchie

Čas (min)	20NewsGroup (10 (dokumen-tov)	20NewsGroup (20 dokumen-tov)	NewYork Times (40 dokumen-tov)	NewYork Times (50 dokumen-tov)
hLDA	**23,48**	**49,33**	**95,13**	**130,20**
HLTA	**0,036**	**0,078**	**0,25**	**0,32**

[1] http://qwone.com/~jason/20Newsgroups/

4 Záver

V práci sme prezentovali prehľad metód z oblasti modelovania témy ako aj jej podoblasti hierarchického modelovania tém, ktoré predstavujú podrobnejšiu analýzu. Práca taktiež zdôrazňovala potrebu a potenciál využitia týchto metód v reálnom svete, a to najmä v oblasti marketingu. Koniec práce bol venovaný porovnaniu metód hLDA a HLTA, kde sa ukázalo, že z hľadiska času spracovania bola metóda LTA oveľa rýchlejšia, avšak z pohľadu kvality vytvorených tém dosahovali tieto metódy porovnateľné výsledky.

Literatúra

1. Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284
2. Hofmann, T. (1999, August). Probabilistic latent semantic indexing. InProceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50-57). ACM.
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.
4. Petterson, J., Buntine, W., Narayanamurthy, S. M., Caetano, T. S., & Smola, A. J. (2010). Word features for latent dirichlet allocation. In Advances in Neural Information Processing Systems (pp. 1921- 1929).
5. Zhai, K., & Boyd-Graber, J. (2013). Online Latent {D} irichlet Allocation with Infinite Vocabulary. In Proceedings of the 30th International Conference on Machine Learning (pp. 561-569).
6. Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2012). Hierarchical dirichlet processes. Journal of the american statistical association.
7. Li, X. M., Ouyang, J. H., & Lu, Y. (2015). Topic modeling for large-scale text data. Frontiers of Information Technology & Electronic Engineering, 16, 457-465.
8. Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. The Journal of Machine Learning Research, 14(1), 1303-1347.
9. Phan, X. H., Nguyen, L. M., & Horiguchi, S. (2008, April). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In Proceedings of the 17th international conference on World Wide Web (pp. 91-100). ACM.
10. Sridhar, V. K. R. (2015, June). Unsupervised topic modeling for short texts using distributed representations of words. In Proceedings of NAACL-HLT (pp. 192-200).
11. Quan, X., Kit, C., Ge, Y., & Pan, S. J. (2015, June). Short and sparse text topic modeling via self-aggregation. In Proceedings of the 24th International Conference on Artificial Intelligence (pp. 2270-2276). AAAI Press.
12. Griffiths, D. M. B. T. L., & Tenenbaum, M. I. J. J. B. (2004). Hierarchical topic models and the nested chinese restaurant process. Advances in neural information processing systems, 16, 17.
13. Hofmann, T. (1999, July). The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. In IJCAI (Vol. 99, pp. 682-687).
14. LI, Wei a Andrew McCALLUM: Pachinko allocation: Scalable micture models of topic correlations. In: Journal of Machine Learning. s.2-30. 2008.

15. CHEN, Peixian a kol.: Progressive EM for latent tree models and hierarchical topic detection. 13th AAAI Conference on Artificial Intelligence. 2016.
16. MCCALLUM, Andrew: Topic model diagnostics [online]. 2002. [cit. 2017-03-15]. Dostupne z: http://mallet.cs.umass.edu/diagnostics.php

Poďakovanie: Tento príspevok vznikol s podporou VEGA projektu č.1/0493/16, KEGA projektu č.025TUKE-4/2015 a APVV projektu č.APVV-16-0213.

Annotation:

Hierarchical topic modeling

Text documents in digital form represent an important source of documents. However currently is their number so large that their manual processing and extraction of information would be time-consuming. For now, there exist several methods of automatic analysis of textual data, one of them is topic modeling. It offers new ways of searching, browsing and summarizing of textual data. For that reason, the main aim of this work is to present methods of topic modeling and also compare selected models for hierarchical topic modeling.