Interval Semi-supervised LDA: Classifying Needles in a Haystack

Svetlana Bodrunova, Sergei Koltsov, Olessia Koltsova,
Sergey Nikolenko, and Anastasia Shimorina

Laboratory for Internet Studies (LINIS),
National Research University Higher School of Economics,
ul. Soyuza Pechatnikov, d. 16, 190008 St. Petersburg, Russia

Abstract. An important text mining problem is to find, in a large collection of texts, documents related to specific topics and then discern further structure among the found texts. This problem is especially important for social sciences, where the purpose is to find the most representative documents for subsequent qualitative interpretation. To solve this problem, we propose an interval semi-supervised LDA approach, in which certain predefined sets of keywords (that define the topics researchers are interested in) are restricted to specific intervals of topic assignments. We present a case study on a Russian LiveJournal dataset aimed at ethnicity discourse analysis.

Keywords: topic modeling, latent Dirichlet allocation, text mining.

1 Introduction

Many applications in social sciences are related to text mining. Researchers often aim to understand how a certain large body of text behaves: what topics interest the authors of this body, how these topics develop and interact, what are the key words that define these topics in the discourse and so on. Topic modeling approaches, usually based on some version of the LDA (latent Dirichlet allocation) model [1], are very important in this regard. Often, the actually interesting part of the dataset is relatively small, although it is still too large to be processed by hand and, moreover, it is unclear how to separate the interesting part from the rest of the dataset. Such an "interesting" part may be, for instance, represented by certain topics that are defined, but not limited to, certain relevant keywords (so that a simple search for these keywords would yield only a subset of the interesting part). In this short paper, we propose a method for identifying documents relevant to a specific set of topics that also extracts its topical structure based on a semi-supervised version of the LDA model. The paper is organized as follows. In Section 2, we briefly review the basic LDA model and survey related work concerning various extensions of the LDA model. In Section 3 we introduce two extensions: semi-supervised LDA that sets a single topic for each predefined set of key words and interval semi-supervised LDA that maps a set of keywords to an interval of topics. In Section 4, we present a case study of mining ethnical

F. Castro, A. Gelbukh, M.G. Mendoza (Eds.): MICAI 2013, Part I, LNAI 8265, pp. 265–274, 2013.

discourse from a dataset of Russian LiveJournal blogs and show the advantages of the proposed approach; Section 5 concludes the paper.

2 The LDA Model and Extensions

2.1 LDA

The basic latent Dirichlet allocation (LDA) model [1,2] is depicted on Fig. 1a. In this model, a collection of D documents is assumed to contain T topics expressed with W different words. Each document $d \in D$ is modeled as a discrete distribution $\theta^{(d)}$ over the set of topics: $p(z_w = j) = \theta^{(d)}$, where z is a discrete variable that defines the topic for each word $w \in d$. Each topic, in turn, corresponds to a multinomial distribution over the words, $p(w \mid z_w = j) = \phi_w^{(j)}$. The model also introduces Dirichlet priors α for the distribution over documents (topic vectors) θ, $\theta \sim \mathrm{Dir}(\alpha)$, and β for the distribution over the topical word distributions, $\phi \sim \mathrm{Dir}(\beta)$. The inference problem in LDA is to find hidden topic variables $\boldsymbol{z}$, a vector spanning all instances of all words in the dataset. There are two approaches to inference in the LDA model: variational approximations and MCMC sampling which in this case is convenient to frame as Gibbs sampling. In this work, we use Gibbs sampling because it generalizes easily to semi-supervised LDA considered below. In the LDA model, Gibbs sampling after easy transformations [2] reduces to the so-called *collapsed Gibbs sampling*, where z_w are iteratively resampled with distributions

$$p(z_w = t \mid \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) \propto$$
$$\propto q(z_w, t, \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) = \frac{n^{(d)}_{-w,t} + \alpha}{\sum_{t' \in T} \left(n^{(d)}_{-w,t'} + \alpha \right)} \frac{n^{(w)}_{-w,t} + \beta}{\sum_{w' \in W} \left(n^{(w')}_{-w,t} + \beta \right)},$$

where $n^{(d)}_{-w,t}$ is the number of times topic t occurs in document d and $n^{(w)}_{-w,t}$ is the number of times word w is generated by topic t, not counting the current value z_w.

2.2 Related Work: LDA Extensions

Over the recent years, the basic LDA model has been subject to many extensions; each of them presenting either a variational of a Gibbs sampling algorithm for a model that builds upon LDA to incorporate some additional information or additional presumed dependencies. Among the most important extensions we can list the following:

– *correlated topic models* (CTM) improve upon the fact that in the base LDA model, topic distributions are independent and uncorrelated, but, of course, some topics are closer to each other and share words with each other; CTM use logistic normal distribution instead of Dirichlet to model correlations between topics [3];

- *Markov topic models* use Markov random fields to model the interactions between topics in different parts of the dataset (different text corpora), connecting a number of different hyperparameters β_i in a Markov random field that lets one subject these hyperparameters to a wide class of prior constraints [4];
- *relational topic models* construct a hierarchical model that reflects the structure of a document network as a graph [5];
- the *Topics over Time* model applies when documents have timestamps of their creation (e.g., news articles); it represents the time when topics arise in continuous time with a beta distribution [6];
- *dynamic topic models* represent the temporal evolution of topics through the evolution of their hyperparameters α and β, either with a state-based discrete model [7] or with a Brownian motion in continuous time [8];
- *supervised LDA* assigns each document with an additional response variable that can be observed; this variable depends on the distribution of topics in the document and can represent, e.g., user response in a recommender system [9];
- *DiscLDA* assumes that each document is assigned with a categorical label and attempts to utilize LDA for mining topic classes related to this classification problem [10];
- the *Author-Topic model* incorporates information about the author of a document, assuming that texts from the same author will be more likely to concentrate on the same topics and will be more likely to share common words [11, 12];
- finally, a lot of work has been done on nonparametric LDA variants based on Dirichlet processes that we will not go into in this paper; for the most important nonparametric approaches to LDA see [13–17] and references therein.

The extension that appears to be closest to the one proposed in this work is the *Topic-in-Set knowledge* model and its extension with Dirichlet forest priors [18,19]. In [19], words are assigned with "z-labels"; a z-label represents the topic this specific word should fall into; in this work, we build upon and extend this model.

3 Semi-supervised LDA and Interval Semi-supervised LDA

3.1 Semi-supervised LDA

In real life text mining applications, it often happens that the entire dataset D deals with a large number of different unrelated topics, while the researcher is actually interested only in a small subset of these topics. In this case, a direct application of the LDA model has important disadvantages. Relevant topics may have too small a presence in the dataset to be detected directly, and one would need a very large number of topics to capture them in an unsupervised fashion.

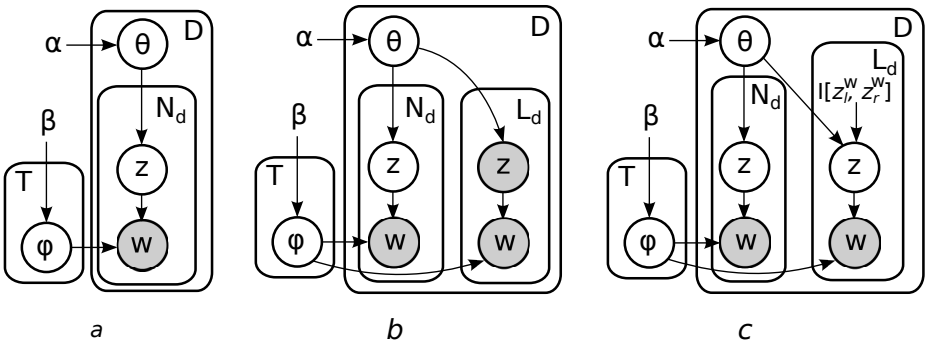

Fig. 1. Probabilistic models: (*a*) LDA; (*b*) semi-supervised LDA; (*c*) interval semi-supervised LDA

For a large number of topics, however, the LDA model often has too many local maxima, giving unstable results with many degenerate topics.

To find relevant subsets of topics in the dataset, we propose to use a semi-supervised approach to LDA, fixing the values of z for certain key words related to the topics in question; similar approaches have been considered in [18, 19]. The resulting graphical model is shown on Fig. 1b. For words $w \in W_{\text{sup}}$ from a predefined set W_{sup}, the values of z are known and remain fixed to $\tilde{z}_w$ throughout the Gibbs sampling process:

$$p(z_w = t \mid \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) \propto \begin{cases} [t = \tilde{z}_w], & w \in W_{\text{sup}}, \\ q(z_w, t, \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) & \text{otherwise.} \end{cases}$$

Otherwise, the Gibbs sampler works as in the basic LDA model; this yields an efficient inference algorithm that does not incur additional computational costs.

3.2 Interval Semi-supervised LDA

One disadvantage of the semi-supervised LDA approach is that it assigns only a single topic to each set of keywords, while in fact there may be more than one topics about them. For instance, in our case study (see Section 4) there are several topics related to Ukraine and Ukrainians in the Russian blogosphere, and artificially drawing them all together with the semi-supervised LDA model would have undesirable consequences: some "Ukrainian" topics would be cut off from the supervised topic and left without Ukrainian keywords because it is more likely for the model to cut off a few words even if they fit well than bring together two very different sets of words under a single topic.

Therefore, we propose to map each set of key words to *several* topics; it is convenient to choose a contiguous interval, hence *interval semi-supervised LDA* (ISLDA). Each key word $w \in W_{\text{sup}}$ is thus mapped to an interval $[z_l^w, z_r^w]$, and the probability distribution is restricted to that interval; the graphical model is

shown on Fig. 1c, where $I[z_l^w, z_r^w]$ denotes the indicator function: $I[z_l^w, z_r^w](z) = 1$ iff $z \in [z_l^w, z_r^w]$. In the Gibbs sampling algorithm, we simply need to set the probabilities of all topics outside $[z_l^w, z_r^w]$ to zero and renormalize the distribution inside:

$$p(z_w = t \mid \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) \propto \begin{cases} I[z_l^w, z_r^w](t) \frac{q(z_w, t, \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta)}{\sum_{z_l^w \le t' \le z_r^w} q(z_w, t', \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta)}, & w \in W_{\text{sup}}, \\ q(z_w, t, \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) & \text{otherwise.} \end{cases}$$

Note that in other applications it may be desirable to assign intersecting subsets of topics to different words, say in a context when some words are more general or have homonyms with other meanings; this is easy to do in the proposed model by assigning a specific subset of topics Z^w to each key word, not necessarily a contiguous interval. The Gibbs sampling algorithm does not change:

$$p(z_w = t \mid \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) \propto \begin{cases} I[Z^w](t) \frac{q(z_w, t, \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta)}{\sum_{t' \in Z^w} q(z_w, t', \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta)}, & w \in W_{\text{sup}}, \\ q(z_w, t, \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) & \text{otherwise.} \end{cases}$$

4 Mining Ethnical Discourse

4.1 Case Study: Project Description

We have applied the method outlined above to a sociological project intended to study ethnical discourse in the Russian blogosphere. The project aims to analyze ethnically-marked discourse, in particular, find: (1) topics of discussion connected to ethnicity and their qualitative discursive interpretation; (2) ethnically-marked social milieus or spaces; (3) ethnically-marked social problems; (4) "just dangerous" ethnicities that would be surrounded by pejorative / stereotyped / fear-marked discourse without any particular reason for it evident from data mining. This study stems from constructivist research on inequalities in socio-economical development vs. ethnical diversity, ethnicities as social borders, and ethnicities as sources of moral panic [20–22]; our project goes in line with current research in mediated representations of ethnicity and ethnic-marked discourse [23–26].

The major issues in mediated representation of ethnicity may be conceptualized as criminalization of ethnicity, sensibilization of cultural difference and enhancing cultural stereotypes, problematization of immigration, reinforcement of negativism in image of ethnicities, unequal coverage of ethnic groups, labeling and boundary marking, and flawed connections of ethnicity with other major areas of social cleavages, e.g. religion. The approach needs to be both quantitative and qualitative; we need to be able to automatically mine existing topics from a large dataset and then qualitatively interpret these results. This led us to topic modeling, and ultimately to developing ISLDA for this project. Thus, the second aim of the project is methodological, as we realize that ethnic vocabulary may not show up as the most important words in the topics, and discursive significance of less frequent ethnonyms (like Tajik or Vietnamese) will be very

low. As for the most frequent ethnonyms (in case of the Russian blogosphere, they include American, Ukranian, or German), our hypothesis was that they may provide varying numbers of discussion topics in the blogosphere, from 0 (no clear topics evident) up to 4 or 5 major topics, which are not always spotted by regular LDA.

4.2 Case Study Results and Discussion

In this section, we present qualitative results of the case study itself and compare LDA with ISLDA. In this case study, the dataset consisted of four months of LiveJournal posts written by 2000 top bloggers. In total, there were 235,407 documents in the dataset, and the dictionary, after cleaning stopwords and low frequency words, contained 192,614 words with about 53.5 million total instances of these words. We have performed experiments with different numbers of topics (50, 100, 200, and 400) for both regular LDA and Interval Semi-Supervised LDA.

Comparing regular LDA results for 100 and 400 topics, it is clear that ethnic topics need to be dug up at 400 rather than 100 topics. The share of ethnic topics was approximately the same: 9 out of 100 (9%) and 34 out of 400 (8.5%), but in terms of quality, the first iteration gives "too thick" topics like Great Patriotic war, Muslim, CEE countries, "big chess play" (great world powers and their roles in local conflicts), Russian vs. Western values, US/UK celebrities and East in travel (Japan, India, China and Korea). This does not provide us with any particular hints on how various ethnicities are treated in the blogosphere.

The 400-topic LDA iteration looks much more informative, providing topics of three kinds: event-oriented (e.g., death of Kim Jong-il or boycotting Russian TV channel NTV in Lithuania), current affairs oriented (e.g., armed conflicts in Libya and Syria or protests in Kazakh city Zhanaozen), and long-term topics. The latter may be divided into "neutral" descriptions of country/historic realities (Japan, China, British Commonwealth countries, ancient Indians etc.), long-term conflict topics (e.g., the Arab-Israeli conflict, Serb-Albanian conflict and the Kosovo problem), and two types of "problematized" topics: internal problems of a given country/nation (e.g., the U.S.) and "Russia vs. another country/region" topics (Poland, Chechnya, Ukraine). There are several topics of particular interest for the ethnic case study: a topic on Tajiks, two opposing topics on Russian nationalism ("patriotic" and "negative"), and a Tatar topic. Several ethnicities, e.g., Americans, Germans, Russians, and Arabs, were subject of more than one topic.

In ISLDA results, the 100-topic modeling covered the same ethnic topics as regular LDA, but Ukrainian ethnonyms produced a new result discussed below. 400-topic ISLDA gave a result much better than regular LDA. For ex-Soviet ethnicities (Tajik and Georgian), one of two pre-assigned topics clearly showed a problematized context. For Tajiks, it was illegal migration: the word collection also showed the writers from opposing opinion camps (Belkovsky, Kholmogorov, Krylov) and vocabulary characteristic of opinion media texts. For Georgians, the context of the Georgian-Ossetian conflict of 2008 clearly showed up, enriched by current events like election issues in South Ossetia. French and Ukrainian, both

assigned 4 topics, showed good results. France had all topics more or less clearly connected to distinctive topics: a Mediterranean topic, Patriotic wars in Russia (with France and Germany), the current conflict in Lybia and general history of Europe. Here, we see that topics related to current affairs are easily de-aligned from long-term topics.

In general, we have found that ISLDA results have significant advantages over regular LDA. Most importantly, ISLDA finds *new important topics* related to the chosen semi-supervised subjects. As an example, Table 1 shows topics from our runs with 100 and 400 topics related to Ukraine. In every case, there is a strong topic related to Ukrainian politics, but then differences begin. In the 100 topic case (Figs. 1a and 1c), ISLDA distinguishes a Ukrainian nationalist topic (very important for our study) that was lost on LDA. With 400 topics (Figs. 1b and 1d), LDA finds virtually the same topics, while ISLDA finds three new important topics: scandals related to Russian natural gas transmitted through Ukraine, a topic devoted to Crimea, and again the nationalist topic (this time with a Western Ukrainian spin). The same pattern appears for other ethnical subjects in the dataset: ISLDA produces more informative topics on the specified subjects.

As for numerical evaluation of modeling results, we have computed the held-out perplexity on two test sets of 1000 documents each; i.e., we estimated the value of

$$p(\mathbf{w} \mid D) = \int p(\mathbf{w} \mid \Phi, \alpha\mathbf{m}) p(\Phi, \alpha\mathbf{m} \mid D) \mathrm{d}\alpha \mathrm{d}\Phi$$

for each held-out document $\mathbf{w}$ and then normalized the result as

$$\text{perplexity}(D_{\text{test}}) = \exp\left(-\frac{\sum_{\mathbf{w}\in D_{\text{test}}} \log p(\mathbf{w})}{\sum_{\mathbf{w}\in D_{\text{test}}} N_d}\right).$$

To compute $p(\mathbf{w} \mid D)$, we used the left-to-right algorithm proposed and recommended in [27, 28]. The test sets were separate datasets of blog posts from the same set of authors and around the same time as the main dataset; the first test set D_{test} contained general posts while the second, $D^{\text{key}}_{\text{test}}$, was comprised of posts that contain at least one of the key words used in ISLDA. Perplexity results are shown in Table 2; it is clear that perplexity virtually does not suffer in ISLDA, and there is no difference in the perplexity between the keyword-containing test set and the general test set. This indicates that ISLDA merely brings the relevant topics to the surface of the model and does not in general interfere with the model's predictive power.

For further sociological studies directed at specific issues, we recommend to use ISLDA with the number of preassigned topics (interval sizes) chosen *a priori* larger than the possible number of relevant topics: in our experiments, we saw that extra slots are simply filled up with some unrelated topics and do not deteriorate the quality of relevant topics. However, the results begin to deteriorate

if more than about 10% of all topics (e.g., 40 out of 400) are assigned to the semi-supervised part; one always needs to have sufficient "free space" to fill with other topics. This provides a certain tension that may be resolved with further study (see below).

Table 1. A comparison of LDA topics related to Ukraine: (*a*) LDA, 100 topics; (*b*) LDA, 400 topics; (*c*) ISLDA, 100 topics; (*d*) ISLDA, 400 topics

(*a*)	Ukraine	0.043	Ukraine	0.049			
	Ukrainian	0.029	Ukrainian	0.017			
	Polish	0.012	Timoshenko	0.015			
	Belorussian	0.011	Yanukovich	0.015			
	Poland	0.011	Victor	0.012			
	Belarus	0.010	president	0.012			
(*b*)	Ukraine	0.098	Ukraine	0.054	dragon	0.026	
	Ukrainian	0.068	Timoshenko	0.019	Kiev	0.022	
	Belorussian	0.020	Yanukovich	0.018	Bali	0.012	
	Belarus	0.018	Ukrainian	0.016	house	0.010	
	Kiev	0.018	president	0.015	place	0.006	
	Kievan	0.012	Victor	0.013	work	0.006	
(*c*)	Ukraine	0.065	Ukraine	0.062	Ukrainian	0.040	Crimea 0.046
	gas	0.030	Timoshenko	0.023	Ukraine	0.036	Crimean 0.015
	Europe	0.026	Ukrainian	0.022	Polish	0.021	Sevastopol 0.015
	Russia	0.019	Yanukovich	0.018	Poland	0.017	Simferopol 0.008
	Ukrainian	0.018	Kiev	0.015	year	0.009	Yalta 0.008
	Belorussian	0.018	Victor	0.014	L'vov	0.006	source 0.007
	Belarus	0.017	president	0.013	Western	0.005	Orjonikidze 0.005
	European	0.015	party	0.013	cossack	0.005	sea 0.005
(*d*)	Ukraine	0.065	Ukraine	0.062	Ukrainian	0.040	Crimea 0.046
	gas	0.030	Timoshenko	0.023	Ukraine	0.036	Crimean 0.015
	Europe	0.026	Ukrainian	0.022	Polish	0.021	Sevastopol 0.015
	Russia	0.019	Yanukovich	0.018	Poland	0.017	Simferopol 0.008
	Ukrainian	0.018	Kiev	0.015	year	0.009	Yalta 0.008
	Belorussian	0.018	Victor	0.014	L'vov	0.006	source 0.007
	Belarus	0.017	president	0.013	Western	0.005	Orjonikidze 0.005
	European	0.015	party	0.013	cossack	0.005	sea 0.005

Table 2. Held-out perplexity results

# of topics	Perplexity, LDA		Perplexity, ISLDA	
	D_{test}	$D_{\text{test}}^{\text{key}}$	D_{test}	$D_{\text{test}}^{\text{key}}$
100	12.7483	12.7483	12.7542	12.7542
200	12.7457	12.7457	12.7485	12.7486
400	12.6171	12.6172	12.6216	12.6216

5 Conclusion

In this work, we have introduced the Interval Semi-Supervised LDA model (ISLDA) as a tool for a more detailed analysis of a specific set of topics inside a larger dataset and have showed an inference algorithm for this model based on collapsed Gibbs sampling. With this tool, we have described a case study in ethnical discourse analysis on a dataset comprised of the Russian LiveJournal blogs. We show that topics relevant to the subject of study do indeed improve in the ISLDA analysis and recommend ISLDA for further use in sociological studies of the blogosphere.

For further work, note that the approach outlined above requires the user to specify how many topics are assigned to each keyword. We have mentioned that there is a tradeoff between possibly losing interesting topics and breaking the model up by assigning too many topics in the semi-supervised part; in the current model, we can only advise to experiment until a suitable number of semi-supervised topics is found. Therefore, we propose an interesting open problem: develop a nonparametric model that chooses the number of topics in each semi-supervised cluster of topics separately and also chooses separately the rest of the topics in the model.

Acknowledgements. This work was done at the Laboratory for Internet Studies, National Research University Higher School of Economics (NRU HSE), Russia, and partially supported by the Basic Research Program of NRU HSE. The work of Sergey Nikolenko was also supported by the Russian Foundation for Basic Research grant 12-01-00450-a and the Russian Presidential Grant Programme for Young Ph.D.'s, grant no. MK-6628.2012.1.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research 3(4-5), 993–1022 (2003)
2. Griffiths, T., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences 101 (suppl. 1), 5228–5335 (2004)
3. Blei, D.M., Lafferty, J.D.: Correlated topic models. Advances in Neural Information Processing Systems 18 (2006)
4. Li, S.Z.: Markov Random Field Modeling in Image Analysis. Advances in Pattern Recognition. Springer (2009)
5. Chang, J., Blei, D.M.: Hierarchical relational models for document networks. Annals of Applied Statistics 4(1), 124–150 (2010)
6. Wang, X., McCallum, A.: Topics over time: a non-Markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 424–433. ACM, New York (2006)
7. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 113–120. ACM, New York (2006)
8. Wang, C., Blei, D.M., Heckerman, D.: Continuous time dynamic topic models. In: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (2008)

9. Blei, D.M., McAuliffe, J.D.: Supervised topic models. Advances in Neural Information Processing Systems 22 (2007)
10. Lacoste-Julien, S., Sha, F., Jordan, M.I.: DiscLDA: Discriminative learning for dimensionality reduction and classification. In: Advances in Neural Information Processing Systems, vol. 20 (2008)
11. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pp. 487–494. AUAI Press, Arlington (2004)
12. Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., Steyvers, M.: Learning author-topic models from text corpora. ACM Trans. Inf. Syst. 28, 1–38 (2010)
13. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. Journal of the American Statistical Association 101(476), 1566–1581 (2004)
14. Blei, D.M., Jordan, M.I., Griffiths, T.L., Tennenbaum, J.B.: Hierarchical topic models and the nested chinese restaurant process. Advances in Neural Information Processing Systems 13 (2004)
15. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Sharing clusters among related groups: Hierarchical Dirichlet processes. Advances in Neural Information Processing Systems 17, 1385–1392 (2005)
16. Williamson, S., Wang, C., Heller, K.A., Blei, D.M.: The IBP compound Dirichlet process and its application to focused topic modeling. In: Proceedings of the 27th International Conference on Machine Learning, pp. 1151–1158 (2010)
17. Chen, X., Zhou, M., Carin, L.: The contextual focused topic model. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 96–104. ACM, New York (2012)
18. Andrzejewski, D., Zhu, X., Craven, M.: Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In: Proc. 26th Annual International Conference on Machine Learning, ICML 2009, pp. 25–32. ACM, New York (2009)
19. Andrzejewski, D., Zhu, X.: Latent Dirichlet allocation with topic-in-set knowledge. In: Proc. NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, SemiSupLearn 2009, pp. 43–48. Association for Computational Linguistics, Stroudsburg (2009)
20. Barth, F.: Introduction. In: Barth, F. (ed.) Ethnic Groups and Boundaries: The Social Organization of Culture Difference, pp. 9–38. George Allen and Unwin, London (1969)
21. Hechter, M.: Internal colonialism: the Celtic fringe in British national development, pp. 1536–1966. Routledge & Kegan Paul, London (1975)
22. Hall, S.: Ethnicity: Identity and difference. Radical America 23(4), 9–22 (1991)
23. Voltmer, K.: The Media in Transitional Democracies. Polity, Cambridge (2013)
24. Nyamnjoh, F.B.: Africa's Media, Democracy and the Politics of Belonging. Zed Books, London (2005)
25. ter Wal, J. (ed.): Racism and cultural diversity in the mass media: An overview of research and examples of good practice in the EU member states, 1995-2000, pp. 1995–2000. European Monitoring Centre on Racism and Xenofobia, Vienna (2002)
26. Downing, J.D.H., Husbands, C.: Representing Race: Racisms, Ethnicity and the Media. Sage, London (2005)
27. Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D.: Evaluation methods for topic models. In: Proceedings of the 26th International Conference on Machine Learning, pp. 1105–1112. ACM, New York (2009)
28. Wallach, H.M.: Structured topic models for language. PhD thesis, University of Cambridge (2008)