Latent Dirichlet Allocation literature review

github:LDA.bib/CREDITS.txt

December 26, 2012

Abstract

We review some of the recent results about topic models.

keywords: LDA, topic model, graphical model.

1 Introduction

[ABFX08, AZC09, ASW08, AWST09, BJ03, Ble03, BNJ03, BGJT03, BGJ07, BL06, BL07, BM07, BGBZ07, BGB07, BGB08, BGB09, Bun09a, BJ05, Bun09b, CLT07, CB09, CBGW$^+$09, III09, DDF$^+$90, DBS07, DLP08, DE09, EFL04, e09, GB10, GK03, GSH10, GHSS09, GS04, GSBT04, HJM08, Hei04, HGG07, Hof99, HBB10, JI10, Joh10, KSJ07, LJSJ08, LD97, LBM07, LXH08, McC02, MCEW05, MLW$^+$07, MSZ07, MCZZ08, MM07a, MM07b, MLM07, MM08, MWM08, MWN$^+$09, MB08, MB08, NAXC08, NCD$^+$07, NCL07, NCS06, NB06, NLGB10, NSHC09, PN07, PBP04, PKGT06, RR09, RHNM09, RDL10, RWSM10, RZGSS04, SH09, SN10, SG05, SG06, TNW06, TJBB06, TJ07, Wal06, WMSM09, WMM09, WMM05, WM06, WBH08, WBL09, LM06, WC06, YXQ09, YMM09, ZX06, ZX07, ZAX09, ZX10, ZBL06]

References

[ABFX08] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *JMLR*, 9:1981–2014, 2008.

[ASW08] Arthur Asuncion, Padhraic Smyth, and Max Welling. Asynchronous distributed learning of topic models. In *NIPS*, pages 81–88, 2008.

[AWST09] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee-Whye Teh. On smoothing and inference for topic models. In *UAI*, 2009. A dense but excellent review of inference in topic models. Introduces CVB0, a method for collapsed variational inference surprisingly similar to Gibbs sampling.

[AZC09] David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *ICML*, pages 25–32, 2009.

[BGB07] Jordan Boyd-Graber and David M. Blei. Putop: Turning predominant senses into a topic model for wsd. In *SEMEVAL*. Association for Computational Linguistics, 2007.

[BGB08] Jordan Boyd-Graber and David M. Blei. Syntactic topic models. In *NIPS*, 2008.

[BGB09] Jordan Boyd-Graber and David M. Blei. Multilingual topic models for unaligned text. In *UAI*, 2009.

[BGBZ07] Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. A topic model for word sense disambiguation. In *EMNLP*, 2007.

[BGJ07] David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested chinese restaurant process and hierarchical topic models, 2007. This is a longer version of.

[BGJT03] David M. Blei, Thomas Griffiths, Michael Jordan, and Joshua Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, 2003. Introduces hLDA, which models topics in a tree. Each document is generated by topics along a single path through the tree.

[BJ03] David Blei and Michael Jordan. Modeling annotated data. In *SIGIR*, 2003. This paper introduces CorrLDA for data that consists of text and images, where image "topics" are chosen only from topics that are assigned to the text in the same document.

[BJ05] Wray L. Buntine and Aleks Jakulin. Discrete component analysis. In *SLSFS*, pages 1–33, 2005.

[BL06] David M. Blei and John D. Lafferty. Dynamic topic models. In *ICML*, 2006.

[BL07] David M. Blei and John D. Lafferty. A correlated topic model of. *AAS*, 1(1):17–35, 2007.

[Ble03] David M. Blei. lda-c, 2003. lda-c implements LDA with variational inference in C.

[BM07] David M. Blei and Jon D. McAuliffe. Supervised topic models. In *NIPS*, 2007.

[BNJ03] David M. Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, January 2003.

[Bun09a] Wray L. Buntine. Discrete component analysis, 2009. C implementation of LDA and multinomial PCA.

[Bun09b] Wray L. Buntine. Estimating likelihoods for topic models. In *Asian Conference on Machine Learning*, 2009. Provides improved versions of some of the methods in.

[CB09] Jonathan Chang and David Blei. Relational topic models for document networks. In *AISTATS*, 2009.

[CBGW+09] Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *NIPS*, 2009.

[CLT07] Jun Fu Cai, Wee Sun Lee, and Yee Whye Teh. Nus-ml: Improving word sense disambiguation using topic features. In *SEMEVAL*. Association for Computational Linguistics, 2007.

[DBS07] Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised prediction of citation influences. In *ICML*, 2007.

[DDF+90] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[DE09] Gabriel Doyle and Charles Elkan. Accounting for burstiness in topic models. In *ICML*, 2009. Replaces the standard multinomial distribution over topics with a Dirichlet-compound Multinomial (DCM).

[DLP08] Chris Ding, Tao Li, and Wei Peng. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics and Data Analysis*, 52:3913–3927, 2008.

[EFL04] Elena Erosheva, Stephen Fienberg, and John Lafferty. Mixed membership models of scientific publications. *PNAS*, 101(Suppl. 1):5220–5227, 2004.

[GB10] Sean Gerrish and David M. Blei. A language-based approach to measuring scholarly impact. In *ICML*, 2010.

[GHSS09] Andre Gohr, Alexander Hinneburg, Rene Schult, and Myra Spiliopoulou. Topic evolution in a stream of documents. In *SDM*, pages 859–870, 2009.

[GK03] Mark Girolami and Ata Kabn. On an equivalence between plsi and lda. In *SIGIR*, pages 433–434, New York, NY, USA, 2003. ACM.

[GS04] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[GSBT04] Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. Integrating topics and syntax. In *NIPS*, pages 537–544. 2004.

[GSH10] Andre Gohr, Myra Spiliopoulou, and Alexander Hinneburg. Visually summarizing the evolution of documents under a social tag. In *KDIR*, 2010.

[HBB10] Matthew Hoffman, David M. Blei, and Francis Bach. Online learning for latent dirichlet allocation. In *NIPS*, 2010.

[Hei04] Gregor Heinrich. Parameter estimation for text analysis. Technical report, 2004.

[HGG07] Alexander Hinneburg, Hans-Henning Gabriel, and Andre Gohr. Bayesian folding-in with dirichlet kernels for plsi. In *ICDM*, pages 499–504, 2007.

[HJM08] David Hall, Daniel Jurafsky, and Christopher D. Manning. Studying the history of ideas using topic models. In *EMNLP*, pages 363–371, 2008.

[Hof99] Thomas Hofmann. Probilistic latent semantic analysis. In *UAI*, 1999.

[III09] Hal Daum III. Markov random topic fields. 2009.

[JI10] Jagadeesh Jagarlamudi and Hal Daum III. Extracting multilingual topics from unaligned comparable corpora. pages 444–456, 2010.

[Joh10] Mark Johnson. Pcfgs, topic models, adaptor grammars, and learning topical collocations and the structure of proper names. 2010.

[KSJ07] Jyri J. Kivinen, Erik B. Sudderth, and Michael I. Jordan. Learning multiscale representations of natural scenes using dirichlet processes. In *ICCV*, 2007. The paper introduces a blocked Gibbs sampler for learning a nonparametric Bayesian topic model whose topic assignments are coupled with a tree-structured graphical model.

[LBM07] Wei Li, David Blei, and Andrew McCallum. Nonparametric bayes pachinko allocation. Technical report, 2007.

[LD97] Thomas K. Landauer and Susan T. Dumais. Solutions to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, (104), 1997.

[LJSJ08] Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *NIPS*, 2008.

[LM06] Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *ICML*, 2006.

[LXH08] Wei-Hao Lin, Eric Xing, and Alexander Hauptmann. A joint topic and perspective model for ideological discourse. In *ECML PKDD*,

pages 17–32, Berlin, Heidelberg, 2008. Springer-Verlag.

[MB08] Indraneel Mukherjee and David Blei. Relative performance guarantees for approximate inference in latent dirichlet allocation. In *NIPS*, 2008.

[McC02] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit, 2002. Implements Gibbs sampling for LDA in Java using fast sampling methods from.

[MCEW05] Andrew McCallum, Andrs Corrada-Emmanuel, and Xuerui Wang. Topic and role discovery in social networks. In *IJCAI*, 2005.

[MCZZ08] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. In *WWW*, 2008.

[MLM07] David Mimno, Wei Li, and Andrew McCallum. Mixtures of hierarchical topics with pachinko allocation. In *ICML*, 2007.

[MLW+07] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW*, 2007.

[MM07a] David Mimno and Andrew McCallum. Expertise modeling for matching papers with reviewers. In *KDD*, 2007.

[MM07b] David Mimno and Andrew McCallum. Mining a digital library for influential authors. In *JCDL*, 2007.

[MM08] David Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI*, 2008. Per-document Dirichlet priors over topic distributions are generated using a log-linear combination of observed document features and learned feature-topic parameters. Implemented in.

[MSZ07] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In *KDD*, pages 490–499, New York, NY, USA, 2007. ACM.

[MWM08] David Mimno, Hanna Wallach, and Andrew McCallum. Gibbs sampling for logistic normal topic models with graph-based priors. In *NIPS Workshop on Analyzing Graphs*, 2008. Introduces an auxiliary-variable method for Gibbs sampling in non-conjugate topic models.

[MWN+09] David Mimno, Hanna Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *EMNLP*, 2009.

[NAXC08] Ramesh Nallapati, Amr Ahmed, Eric P. Xing, and William Cohen. Joint latent topic models for text and citations. In *KDD*, pages 542–550, 2008. This is one of the first papers to address joint topic models of text and hyperlinks. Used as a baseline in the more recent Relational Topic Models. (R.N.).

[NB06] D. Newman and S. Block. Probabilistic topic decomposition of an eighteenth-century american newspaper. *JASIST*, 2006.

[NCD+07] Ramesh Nallapati, William Cohen, Susan Ditmore, John Lafferty, and Kin Ung. Multi-scale topic tomography. In *KDD*, pages 520–529, 2007. Models variation of topic content with time at various scales of resolution. A novel variant of dynamic topic models that uses the Poisson distribution for word generation, and wavelets. (R.N.).

[NCL07] Ramesh Nallapati, William Cohen, and John Lafferty. Parallelized variational em for latent dirichlet allocation: An experimental evaluation of speed and scalability. In *ICDM workshop on high performance data mining*, 2007. Early paper on parallel implementations of variational EM for LDA. (R.N.).

[NCS06] David Newman, Chaitanya Chemudugunta, and Padhraic Smyth. Statistical entity-topic models. In *KDD*, 2006.

[NLGB10] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *NAACL*, 2010.

[NSHC09] Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. Mining multilingual topics from wikipedia. In *WWW*, 2009.

[PBP04] Jukka Perki, Wray L. Buntine, and Sami Perttu. Exploring independent trends in a topic-based search engine. In *Web Intelligence*, pages 664–668, 2004.

[PKGT06] Matthew Purver, Konrad Krding, Thomas L. Griffiths, and Joshua Tenenbaum. Unsupervised topic modelling for multi-party spoken discourse. In *ACL*, 2006.

[PN07] Xuan-Hieu Phan and Cam-Tu Nguyen. Gibbslda++, 2007. C/C++ implementation of LDA with Gibbs sampling.

[RDL10] Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.

[RHNM09] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, 2009.

[RR09] Daniel Ramage and Evan Rosen. Stanford topic modeling toolbox, 2009. Scala implementation of LDA and.

[RWSM10] Joseph Reisinger, Austin Waters, Brian Silverthorn, and Raymond J. Mooney. Spherical topic models. In *ICML*, 2010.

[RZGSS04] Michal Rosen-Zvi, Tom Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *UAI*, 2004.

[SG05] Mark Steyvers and Tom Griffiths. Matlab topic modeling toolbox, 2005. Implements LDA, Author-Topic, HMM-LDA, LDA-COL. Tools for 2D visualization.

[SG06] Mark Steyvers and Tom Griffiths. Probabilistic topic models. In T. Landauer, editor, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2006.

[SH09] Ruslan Salakhutdinov and Geoffrey Hinton. Replicated softmax: an undirected topic model. In *NIPS*, 2009.

[SN10] Alexander Smola and Shravan Narayanamurthy. An architecture for parallel topic models. In *VLDB*, 2010.

[TJ07] Kristina Toutanova and Mark Johnson. A bayesian lda-based model for semi-supervised part-of-speech tagging. In *NIPS*, pages 1521–1528, 2007.

[TJBB06] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *JASA*, 101, 2006.

[TNW06] Yee-Whye Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *NIPS*, 2006.

[Wal06] Hanna M. Wallach. Topic modeling: beyond bag-of-words. In *ICML*, 2006.

[WBH08] Chong Wang, David M. Blei, and David Heckerman. Continuous time dynamic topic models. In *UAI*, 2008.

[WBL09] Chong Wang, David Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *CVPR*, 2009.

[WC06] Xing Wei and Bruce Croft. Lda-based document models for ad-hoc retrieval. In *SIGIR*, 2006.

[WM06] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *KDD*, 2006.

[WMM05] Xuerui Wang, Natasha Mohanty, and Andrew McCallum. Group and topic discovery from relations and their attributes. In *NIPS*, 2005.

[WMM09] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In *NIPS*, 2009. The use of an asymmetric Dirichlet prior on per-document topic distributions reduces sensitivity to very common words (eg stopwords and near-stopwords) and makes topic assignments more stable as the number of topics grows.

[WMSM09] Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *ICML*, 2009. Commonly used methods for estimating the probability of held-out words may be unstable. This paper presents more accurate methods.

[YMM09] Limin Yao, David Mimno, and Andrew McCallum. Efficient methods for topic model inference on streaming document collections. In *KDD*, 2009. Explores methods for inferring topic distributions for new documents given a trained model. This paper includes the SparseLDA algorithm and data structure, which can dramatically improve time and memory performance in Gibbs sampling.

[YXQ09] Feng Yan, Ningyi Xu, and Yuan Qi. Parallel inference for latent dirichlet allocation on graphics processing units. In *NIPS*, 2009. In addition to dividing the corpus between processors, this work divides the vocabulary into the same number of partitions, such that each processor works on both its own documents and its own words at each epoch. This increases the number of epochs, but drastically reduces the possibility of incorrect samples.

[ZAX09] Jun Zhu, Amr Ahmed, and Eric P. Xing. Medlda: Maximum margin supervised topic models for regression and classification. In *ICML*, 2009.

[ZBL06] Xiaojin Zhu, David M. Blei, and John Lafferty. Taglda: Bringing document structure knowledge into topic models. Technical Report TR-1553, University of Wisconsin, Madison, 2006.

[ZX06] Bing Zhao and Eric P. Xing. Bitam: Bilingual topic admixture models for word alignment. In *ACL*, 2006.

[ZX07] Bin Zhao and Eric P. Xing. Hm-bitam: Bilingual topic exploration, word alignment, and translation. In *NIPS*, 2007.

[ZX10] Jun Zhu and Eric P. Xing. Conditional topic random fields. In *ICML*, 2010.

[e09] Radim ehek. gensim, 2009. Python package for topic modelling, includes distributed and online implementation of variational LDA.