中国科学院文献情报中心主办

现代图书情报技术

NEW TECHNOLOGY OF LIBRARY AND INFORMATION SERVICE

本刊检索 GO

高级检索 检索说明

首 页 | 关于我们 | 投稿指南 | 征订服务 | 诚邀合作 | 留 言

现代图书情报技术 » 2012, Vol. » Issue (12) :58-65

情报分析与研究 最新目录 | 下期目录 | 过刊浏览 | 高级检索 << Previous Articles | Next Articles >>

利用LDA的领域新兴主题探测技术综述

范云满1,2, 马建霞1

1. 中国科学院国家科学图书馆兰州分馆 兰州 730000;
2. 中国科学院大学 北京 100049

Fan Yunman1,2, Ma Jianxia1

1. The Lanzhou Branch of National Science Library, Chinese Academy of Sciences, Lanzhou 730000, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China

- 摘要
- 参考文献
- 相关文章

Download: PDF (1147KB) HTML (KB) Export: BibTeX or EndNote (RIS) Supporting Info

摘要 以LDA为基础,系统梳理新兴主题探测以及主题趋势探测技术中的LDA以及其他LDA改进主题模型的发展现状。介绍LDA的变分推导和Gibbs抽样两种参数推导算法;总结近年来LDA模型的改进,包括对主题演化建模的主题模型、对文档内容和元数据联合建模的模型、采用在线式学习的主题模型及将LDA和引文分析相结合的主题演化方法等,并对不同的改进模型进行深入对比和分析;梳理NIH-VB、TIARA、VxInsight等几种主要的主题模型可视化技术。最后通过对LDA模型的总结分析,探讨利用LDA模型探测领域新兴主题时的关键研究问题。

关键词: 主题模型 LDA 引文分析 主题模型可视化

Abstract: Based on LDA,this paper reviews the development of the LDA model and several models which improve the LDA for the filed emerging topic detection.It describes two parameter inference algorithms of variational derivation and Gibbs sampling, and reviews the improvement of LDA in recent years,including the one modeling the evolution of the topics,the one modeling jointly with the content of document and meta data,the one with online learning, the topic evolution method combining LDA and citation analysis and so on;then compares and analyses different kinds of improvement models in details. The paper also reviews several main visualization techniques such as NIH-VB,TIARA and VxInsight. Finally,it discusses the key research problems of detecting the emerging topic by using LDA.

Keywords: Topic model, LDA, Citation analysis, Topical visualization

收稿日期: 2012-10-15;

基金资助:本文系中国科学院西部之光联合学者基金项目“基于计算情报方法的甘肃省战略新兴产业技术创新竞争与发展研究”的研究成果之一。

通讯作者 范云满 Email: fanyunman@mail.las.ac.cn

Service

把本文推荐给朋友

加入我的书架

加入引用管理器

Email Alert

RSS

作者相关文章

范云满

马建霞

引用本文:

范云满, 马建霞 .利用LDA的领域新兴主题探测技术综述[J] 现代图书情报技术, 2012,V(12): 58-65

Fan Yunman, Ma Jianxia .Review on the LDA-based Techniques Detection for the Field Emerging Topic[J] , 2012,V(12): 58-65

链接本文:

http://www.infotech.ac.cn/CN/ 或 http://www.infotech.ac.cn/CN/Y2012/V/I12/58

[1] Blei D M. Probabilistic Topic Models[J]. *Communications of the ACM*, 2012, 55(4): 77-84.

[2] Nigam K, Mccallum A K, Thrun S, et al. Text Classification from Labeled and Unlabeled Documents Using EM[J]. *Machine Learning*, 2000, 39 (2-3): 103-134.

[3] Hofmann T. Probabilistic Latent Semantic Indexing[C]. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*. New York: ACM, 1999: 50-57.

[4] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. *Journal of Machine Learning Research*, 2003, 3: 993-1022.

[5] Jordan M I, Ghahramani Z, Jaakkola T S, et al. An Introduction to Variational Methods for Graphical Models[J]. *Machine learning*, 1999, 37 (2): 183-233.

[6] Teh Y W, Newman D, Welling M. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation[C]. In: *Proceedings of*

Neural Information Processing Systems. 2006: 1353-1360.

[7] Griffiths T. Gibbs Sampling in the Generative Model of Latent Dirichlet Allocation[OL]. [2012-06-09].http://people.cs.umass.edu/~wallach/courses/s11/cmpsci791ss/readings/griffiths02gibbs.pdf.

[8] Heinrich G. Parameter Estimation for Text Analysis[EB/OL]. [2012-06-09]. http://www. arbylon. net/publications/text-est. pdf.

[9] Wainwright M J, Jordan M I. Graphical Models, Exponential Families, and Variational Inference[J]. *Foundations and Trends in Machine Learning*, 2008,1 (1-2): 1-305.

[10] Ghahramani Z, Beal M J. Graphical Models and Variational Methods[A]. //Advanced Mean Field Methods:Theory and Practice[M]. Cambridge: MIT Press, 2001: 167-177.

[11] Blei D M, Lafferty J D. A Correlated Topic Model of Science[J]. *Annals of Applied Statistics*, 2007, 1(1):17-35.

[12] Aldous D J. Exchangeability and Related Topics[M].Berlin, Heidelberg: Springer, 1985: 1-198.

[13] Li W, Mccallum A. Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations[C]. In: *Proceedings of the 23rd International Conference on Machine Learning (ICML ’06)*. New York: ACM, 2006: 577-584.

[14] Wang C, Blei D M. A Split-Merge MCMC Algorithm for the Hierarchical Dirichlet Process[J/OL]. *Computing Research Repository*. [2012-09-24]. http://arxiv.org/abs/1201.1657.

[15] 曹娟,张勇东,李锦涛,等. 一种基于密度的自适应最优LDA模型选择方法[J]. 计算机学报, 2008, 31(10): 1780-1787. (Cao Juan, Zhang Yongdong, Li Jintao, et al. A Method of Adaptively Selecting Best LDA Model Based on Density[J]. *Chinese Journal of Computers*, 2008, 31(10): 1780-1787.)

[16] Blei D M, Lafferty J D. Dynamic Topic Models[C]. In: *Proceedings of the 23rd International Conference on Machine Learning (ICML ’06)*. New York: ACM, 2006: 113-120.

[17] Wang C, Blei D M, Heckerman D. Continuous Time Dynamic Topic Models[C]. In: *Proceedings of Uncertainty in Artificial Intelligence*. 2008: 579-586.

[18] Wang X R, McCallum A. Topics Over Time: A Non-Markov Continuous-time Model of Topical Trends[C]. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’06)*. New York: ACM, 2006: 424-433.

[19] Wallach H M. Topic Modeling: Beyond Bag-of-words[C]. In: *Proceedings of the 23rd International Conference on Machine Learning (ICML ’06)*. New York: ACM, 2006: 977-984.

[20] Wang X R, McCallum A, Wei X. Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval[C]. In: *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM ’07)*. Washington, DC: IEEE Computer Society, 2007: 697-702.

[21] Wang X R, McCallum A. A Note onTopical N-grams[R]. 2005.

[22] Mann G S, Mimno D, McCallum A. Bibliometric Impact Measures Leveraging Topic Analysis[C]. In: *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL ’06)*. New York: ACM, 2006: 65-74.

[23] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The Author-topic Model for Authors and Documents[C]. In: *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI ’04)*. Arlington: AUAI Press, 2004: 487-494.

[24] 王萍. 基于概率主题模型的文献知识挖掘[J]. 情报学报, 2011, 30(6): 583-590. (Wang Ping. Literature Knowledge Mining Based on Probabilistic Topic Model[J]. *Journal of the China Society for Scientific and Technical Information*, 2011, 30(6): 583-590.)

[25] Mimno D, McCallum A. Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression[C]. In: *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence (UAI ’08)*. 2008: 411-418.

[26] Nallapati R M, Ahmed A, Xing E P, et al. Joint Latent Topic Models for Text and Citations[C]. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’08)*. New York: ACM, 2008: 542-550.

[27] Tu Y N, Seng J L. Indices of Novelty for Emerging Topic Detection[J]. *Information Processing & Management*, 2012, 48(2): 303-325.

[28] Goodrum A A, McCain K W, Lawrence S, et al. Scholarly Publishing in the Internet Age: A Citation Analysis of Computer Science Literature [J]. *Information Processing & Management*, 2001, 37(5): 661-675. Information Processing target="_blank">

[29] Web of Knowledge [DB/OL]. [2012-08-14]. http://apps.webofknowledge.com.

[30] 中华人民共和国国家知识产权局.专利检索[EB/OL]. [2012-08-14]. http://www.sipo.gov.cn/zljs/. (State Intellectual Property Office of PRC. Patent Retrieval[EB/OL]. [2012-08-14]. http://www.sipo.gov.cn/zljs/.)

[31] Dietz L, Bickel S, Scheffer T. Unsupervised Prediction of Citation Influences[C]. In: *Proceedings of the 24th International Conference on Machine Learning (ICML ’07)*. New York: ACM, 2007: 233-240.

[32] He Q, Chen B, Pei J, et al. Detecting Topic Evolution in Scientific Literature: How Can Citations Help[C]. In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM ’09)*. New York: ACM, 2009: 957-966.

[33] 贺亮, 李芳. 基于话题模型的科技文献话题发现和趋势分析[J]. 中文信息学报, 2012, 26(2): 109-115.(He Liang, Li Fang. Topic Discovery and Trend Analysis in Scientific Literature on Topic Model[J]. *Journal of Chinese Information Processing*, 2012, 26(2): 109-115.)

[34] Alsumait L, Barbará D, Domeniconi C. On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking[C]. In: *Proceedings of the 8th IEEE International Conference on Data Mining*. 2008: 3-12.

[35] Hoffman M D, Blei D M, Bach F. Online Learning for Latent Dirichlet Allocation[A]. //Lafferty J,Williams C K I,Shawe-Taylor J,et al. Advances in Neural Information Processing Systems[M].2010: 856-864.

[36] Banerjee A, Basu S. Topic Models over Text Streams: A Study of Batch and Online Unsupervised Learning[C]. In: *Proceedings of SDM-SIAM International Conference on Data Mining*. 2007.

[37] Herr B W, Talley E M, Burns G, et al. The NIH Visual Browser: An Interactive Visualization of Biomedical Research[C]. In: *Proceedings of the 13th International Conference Information Visualization (IV ' 09)*. Washington D C: IEEE Computer Society, 2009: 505-509.

[38] Talley E M, Newman D, Mimno D, et al. Database of NIH Grants Using Machine-learned Categories and Graphical Clustering[J]. *Nature Methods*, 2011, 8(6): 443-444.

[39] Wei F R, Liu S X, Song Y Q, et al. TIARA: A Visual Exploratory Text Analytic System[C]. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ' 10)*, Washington DC, USA. New York: ACM, 2010: 153-162.

[40] Boyack K W, Wylie B N, Davidson G S. Domain Visualization Using VxInsight? For Science and Technology Management[J]. *Journal of the American Society for Information Science and Technology*, 2002, 53 (9): 764-774.

[1] 陈仕吉, 史丽文, 李冬梅, 左文革.论文被引频次标准化方法述评[J]. 现代图书情报技术, 2012,28(4): 54-60

[2] 胡志刚, 陈超美, 刘则渊, 侯海燕.基于XML全文数据引文分析系统的设计与实现[J]. 现代图书情报技术, 2012,(11): 72-77

[3] 单斌, 李芳.基于种子文档LDA话题的演化研究[J]. 现代图书情报技术, 2011,27(7/8): 104-109

[4] 李欣, 刘丹.基于LDAP实现多认证源的统一身份认证实践——以华东师范大学图书馆为例[J]. 现代图书情报技术, 2011,27(4): 89-93

[5] 陈仕吉.科学研究前沿探测方法综述[J]. 现代图书情报技术, 2009,(9): 28-33

[6] 陈亦佳,赵星.基于期刊引文网络视角研究国际图书馆学情报学知识交流[J]. 现代图书情报技术, 2009,25(6): 55-60

[7] 孙涛涛,Steven A.Morris,黄亚明.基于专利引文分析的时间线技术[J]. 现代图书情报技术, 2008,24(6): 51-55

[8] 刘佳佳,董茗,方曙 .国外专利分析工具的比较研究[J]. 现代图书情报技术, 2007,2(2): 67-74

[9] 陈祖琴,郑宏 .基于元搜索引擎的中文数据库引文分析系统[J]. 现代图书情报技术, 2006,1(11): 65-68

[10] 许鑫,苏新宁,陆炯.数字化校园身份认证系统的设计[J]. 现代图书情报技术, 2005,21(4): 51-57

[11] 许鑫,苏新宁,姚毅.数字化校园中统一身份认证系统的分析[J]. 现代图书情报技术, 2005,21(3): 50-56

[12] 张智雄.目录及其在分布式主题网关中的应用[J]. 现代图书情报技术, 2003,19(5): 41-44

[13] 王知津,孙美丽.1998-2000年《现代图书情报技术》引文及影响因子分析[J]. 现代图书情报技术, 2003,19(3): 8-11

[14] 程刚.《现代图书情报技术》被引分析[J]. 现代图书情报技术, 2001,17(1): 33-36

[15] 杨新涯.《现代图书情报技术》的文献计量分析[J]. 现代图书情报技术, 1996,12(4): 56-59