推特——网络主题模型：社交网络和文本模型的全贝叶斯处理

Kar Wai Lim
ANU, NICTA
Canberra, Australia

Changyou Chen
ANU, NICTA
Canberra, Australia

Wray Buntine
NICTA, ANU
Canberra, Australia

摘要

推特的数据噪音很多-每个噪音是短的、非结构化非正式的语言，是目前主题建模的挑战。另一方面，噪音伴随着额外的信息，如作者、标签和用户关注网络。利用这些额外的信息，我们提出了推特网络（TN）主题模型，来为文本和社会网络用完整非参数贝叶斯方法进行联合建模。TN 主题模型采用分层的泊松狄利克雷过程（PDP）为文本进行建模，并用高斯过程的随机函数模型对社会网络建模。我们证明了，TN 主题模型明显优于几个现有的非参数模型，因为其灵活性。此外，TN 主题模型可以挖掘额外信息，如作者的兴趣，标签分析，而且引申出进一步的应用，如作者推荐、自动主题标注和标签的建议。注意我们的一般推理框架可以用嵌入式 PDP 节点很容易地应用到其他主题模型中去。

1 引言

博客、微博和社交网站等 Web 服务的出现，让人们可以公开地贡献信息。这个用户生成的信息通常更个人，非正式化，往往包含个人意见。总的来说，它是很有用的，它可以为实体和产品进行信誉分析，检测自然灾害，获得一手消息，甚至人口分析。Twitter，一个易于访问的信息来源，允许用户在叫做噪声发简短文本中表达他们的意见和想法。

LDA 文档主题生成模型（LDA）（Blei et al., 2003）是主题模型的流行形式。不幸的是，LDA 对推特的直接应用产生不良结果，因为文本短，经常有噪声。（Zhao et al., 2011）即微博是非结构化的，往往包含语法和拼写错误，以及非正式的词语如用户定义的缩写，因为有 140 字符的限制。LDA 在短推上失败，因为它是严重依赖词同现。另外值得注意的是，文本消息可能包含特殊标记称为标签；他们作为关键词，让用户自己的推特与其他贴上这个标签的推特进行链接。然而，标签并不正式因为它们没有标准。标签可以被用来作为内联词或分类标签。因此，而不是硬标签，标签最好是作为特殊的词表达推特的主题。因此，推特对于主题模型具有挑战性，因此用点对点模型替代了。在其他的文本分析应用，推特往往通过 NLP 的方法进行清洗，如词汇规范化（Baldwin et al., 2013）。然而，规范化的使用也具有争议（Eisenstein, 2013）。

在本文中，我们提出了一个通过利用伴随推特的辅助信息进行短文本建模的新方法。这一信息，补充词共现，可以让我们更好的对推特进行建模，以及开放更多的应用，如用户推荐和标签建议。我们的主要共现包括：1）一个对推特建模非常好的模型：叫做推特网络主题模型（TN）的完整非参数贝叶斯模型。和 2）结合分层泊松狄利克雷过程（HPDP）和高斯过程（GP）来联合模型的文本、标签、作者和粉丝网。我们还开发了一个任意 PDP 网络灵活的框架，它允许快速部署（包括推理）对 HPDP 主题模型的新变种。尽管 TN 主题模型的复杂度高，但其实施框架使用起来相对简单。

2 背景和相关工作

LDA 是经常用于不同类型的数据，典型的例子有，使用辅助信息的是作者-主题模型（Rosen-Zvi et al., 2004），标签-主题模型（Tsai, 2011），和主题-链接模型（Liu et al., 2009）。

然而，这些模型只处理一种额外的信息，并不能很好地用于推特，因为它们是专用于处理其他类型的文本数据。注意标签-主题模型将标签视为硬标签并使用他们来对文本文件分组，这是不适合由于推特，因为标签噪音的自然性。推特-LDA （Zhao et al., 2011）和行为-主题模型（Qiu et al., 2013）被设计用于对推特明确的建模。两个模型都不是掺和的模型，因为它们对每个文档限制一个标签。行为主题模型分析了用户推荐的每个主题的推特“发帖行为”。另一方面，该 biterm-主题模型（Yan et al., 2013）只使用 biterm 共生模型的推特，丢弃了文档层面的信息。Biterm-主题模型和推特- LDA 没有纳入任何辅助信息。所有上述的主题模式也有限制，因为主题数量需要事先选好，但这是很难的因为我们不知道这个数字。

为了回避选题数量的需要，(Teh and Jordan, 2010)提出的 Hierarchical Dirichlet 法(HDP) LDA，利用了 Dirichlet 过程（DP）作为先验非参数。此外，有一个模型可以用泊松狄利克雷过程取代 DP(PDP，也称为转向你的过程)，它对自然语言中的词频分布建模为幂律模型。在自然语言中，词的频率分布呈现幂律（Goldwater et al., 2006）。对于主题模型，与 PDP 一起取代 Dirichlet 分布可以获得很大的改进（Sato and Nakagawa, 2010）。

最近有很多工作用网络信息对文本数据进行建模（(Liu et al., 2009; Chang and Blei, 2010; Nallapati et al., 2008)），然而，这些模型在本质上是参数化的是有限制性的。相反，Miller 等人（Miller et al., 2009）和 Lloyd 等人（Lloyd et al., 2012）直接用非参数的先验知识对网络数据进行建模，即通过印度自助流程和高斯过程分别做，但不对文本建模。

3 模型概括

TN 的主题模型，利用附带的标签，作者，和粉丝网来更好的为推特建模。TN 主题模型主要有两个主题部分组成：一个对于文本和标签的 HPDP 主题模型，和一个对于粉丝网的基于 GP 的随机映射模型。作者信息用于把两个模型联系在一起。

我们这样设计我们对于文本的 HPDP 主题模型。首先，生成全球主题分布 μ_0，作为一个先验知识。然后为每个作者生成作者的主题分布 ν，和杂项主题分布 μ_1 用于捕捉话题偏离作者的常见主题。给定 ν 和 μ_1，我们为文档和单词生成主题分布(η, θ', θ)。我们还明确地为标签对单词的影响进行建模。在这里不讨论用标准 LDA 进行标签和单词的生成。注意，标签是用单词来表现的，即标签#happy 与单词 happy 分享了相同的令牌。还要注意所有的概率向量分布是通过 PDP 来建模的，使得模型成为网络中的 PDP 节点。

网络建模是通过作者主题分布 ν 连接到的 HPDP 主题模型，在这里，我们把 ν 看做网络模型中 GP 的输入。GP，用 F 标识，决定了作者（x）之间的联系。图 1 展示了 TN 的图示模型，其中区域ⓐ和ⓑ详细地展示了网络模型和主题模型。详细描述见补充材料[1]。我们强调了我们队网络模型的处理放方式与（Lloyd et al., 2012）模型的处理方式是不同的。我们在我们的网络模型中定义了一个新的基于余弦相似度的核心方程，这相对于传统的方程有很大的提升。此外，我们通过加性耦合的主题分布和网络连接，为推理推导出一个新的采样过程。

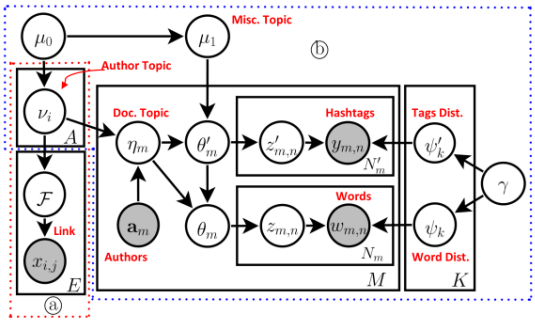

Figure 1: Twitter-Network topic model

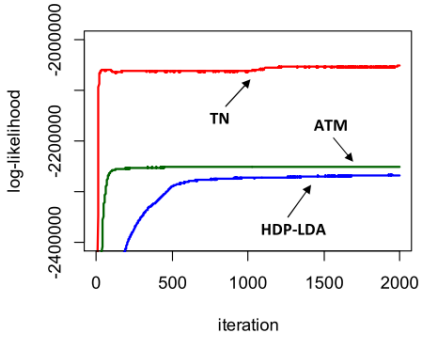

Figure 2: Log-likelihood vs. iterations

[1] Supplementary material is available online at the authors' websites.

4 后验推断

我们选择性地对采样后的主题模型和网络模型执行马尔可夫链 Monte Carlo（MCMC），是互为条件的。我们得到一个主题模型的吉布斯抽样，和网络模型的都市黑斯廷斯（MH）算法。我们开发一个框架来一般在 PDP 的贝叶斯网络上执行，建立在工作（Buntine et al .,2010;Chen et al .,2011），它允许快速构建原型和开发新的主题模型的变体。我们向读者介绍有关技术细节的补充资料。

5 实验和应用

我们用主题模型标准定量评估 TN 主题模型，比如测试集的迷惑度，可能收敛和聚类方法。定性上，我们通过可视化的主题摘要，作者的主题分布，并通过执行自动标记任务来评估我们的模型。我们比较我们的模型与 HDP-LDA 模型，我们是作者主题模型（ATM）的非参数变异，和原来的随机映射网络模型。我们还进行消融研究，以显示模型中的每个组件的重要性。比较和消融研究的结果见表 1。我们使用了两个推特语料进行实验，第一个是推特 7 数据集[2]的一个子集（Yang and Leskovec, 2011），通过用某些关键词进行搜索获得（例如金融， 体育，政治）。我们用 langid.py（Lui and Baldwin, 2012）删除用户发布的不是英语的推特，过滤没有网络信息和少于 100 条推讯的作者。语料库包含 94 个作者的 60370 条推讯。然后我们随机选择 90%的数据集作为训练文档用其他的作为测试文档。第二个语料取自（Mehrotra et al., 2013），共 781186 条推讯。我们注意到，我们没有执行单词归一化，以防止有意义的噪音的文本的任何损失。

Table 1: Perplexity & network log-likelihood

	Perplexity	Network
HDP-LDA	$358.1_{\pm 6.7}$	N/A
ATM	$302.9_{\pm 8.1}$	N/A
Random Function	N/A	$-294.6_{\pm 5.9}$
No Author	$243.8_{\pm 3.4}$	N/A
No Hashtag	$307.5_{\pm 8.3}$	$-269.2_{\pm 9.5}$
No μ_1 node	$221.3_{\pm 3.9}$	$-271.2_{\pm 5.2}$
No Word-tag link	$217.6_{\pm 6.3}$	$-275.0_{\pm 10.1}$
No Power-law	$222.5_{\pm 3.1}$	$-280.8_{\pm 15.4}$
No Network	$218.4_{\pm 4.0}$	N/A
Full TN	$\mathbf{208.4}_{\pm 3.2}$	$\mathbf{-266.0}_{\pm 6.9}$

Table 2: Labeling topics with hashtags

	Top hashtags/words
T0	**#finance #money #economy**
	finance money bank marketwatch stocks china group
T1	**#politics #iranelection #tcot**
	politics iran iranelection tcot tlot topprog obama
T2	**#music #folk #pop**
	music folk monster head pop free indie album gratuit

Table 3: Topics by authors

Twitter ID	Top topics represented by hashtags
finance_yard	#finance #money #realestate
ultimate_music	#music #ultimatemusiclist #mp3
seriouslytech	#technology #web #tech
seriouspolitics	#politics #postrank #news
pr_science	#science #news #postrank

Table 4: Cosine similarity

Recommended	1st	2nd	3rd
Original	0.00	0.05	0.06
TN	**0.78**	**0.57**	**0.55**
Not-recommended	1st	2nd	3rd
Original	0.36	0.33	0.14
TN	**0.17**	**0.09**	**0.10**

实验设置 在所有情况下，我们让 α 在主题节点(μ_0 , μ_1 , ν_i, η_m , θ'_m, θ_m)上，从 0.3 到 0.7 之间变动，在词表节点(ψ, γ)上令 α=0.7 来减弱幂律分布。我们把 β 初始化为 0.5，并设置其先验为 Gamma(0.1, 0.1)。我们设定超参数 λ's, *s*, *l* 和 σ 为 1，因为它们的值在模型结果中没有影响。在下面的评估中，我们用采样算法进行 2000 次迭代来训练概率收敛。我们把每个实验重复做 5 遍来减少评估措施的估计误差。在对 TN 主题模型进行的实验中，我们通过

[2] http://snap.stanford.edu/data/twitter7.html

在整个的推理过程之前首先对吉布斯采样进行 1000 次迭代达到了一个更好的计算效果。在图 2 中，我们可以看到 TN 主题模型相比于 HDP-LDA 和非参数 ATM 收敛速度快。同时，对 TN 主题模型的概率训练在经过 1000 次迭代后可能成为更好的采样网络信息。

自动主题标注 最近有研究致力于在主题建模中自动为主题打标签。在这里，我们表明，使用标签信息使我们能够得到很好的标签作为主题。表 2 显示了通过 TN 主题模型所标注的主题。更详细的主题摘要在补充材料中展示。我们实证评估了代表主题标签的适用性，并发现，总是超过 90%的标签是很好的候选主题标签。

作者主题分布的推断 除了推理在每个文档的主题分布，TN 主题模型使我们能够分析每个作者的主题分布。表 3 展示了不同的作者的主题的概况，在推特 ID 中主题是显而易见的。

作者的建议 我们举例说明作者推荐的 TN 主题模型的使用。在一个新的有 90451 个推文和 625 个新作者的测试数据集中，通过训练 60370 条推文得到的模型，我们预测新作者中最相似和最相似同的作者。我们用作者的主题分布的余弦相似度来量化推荐质量推荐作者对。我们把我们新的核函数和在（Lloyd et al., 2012）中用到的原始的核函数（标记为原始）。表 4 显示了推荐和不推荐作者之间的平均余弦相似度。这表明，我们的核函数更合适。此外，我们手动检查推荐的作者，我们发现，他们通常属于同一个群体，即有类似主题的推文。

聚类和主题连贯性 我们还评价了 TN 主题模型抗衡基于 LAD 的聚类技术（Mehrotra et al., 2013）的艺术状态。我们发现 TN 主题模型在纯度方面优于艺术状态、 归一化互信息和逐点相互信息（PMI）。由于空间限制，评价结果是在补充材料中提供的。

6 结论和将来工作

我们提出了联合为推文和相关的社会网络信息建模的完整的贝叶斯参数推特-网络(TN)主题模型。我们的模型通过 PDP 和 GP，采用非参数的贝叶斯方法对 PDP 网络进行推理，实现了灵活的建模。我们与推特数据集的实验表明，TN 主题模型与现有基线实验相比得到明显改善。此外，我们还研究了 TN 模型的每个组件的可用性。我们的模型也展现了有趣的应用，例如作者建议，以及提供额外的信息推断。

我们还设计了一个框架，用于快速的主题模型开发，由于模型的复杂性，这是非常重要的。虽然我们可以使用适配器语法（Johnson et al., 2007），我们的框架却对主题模型产生更有效的计算。

未来的工作包括加快后验推理算法，特别是对网络模型，以及加强在社交媒体中其他辅助信息，如位置，超链接和多媒体内容。我们还将探索其他的可以用 TN 主题模型解决的应用程序，如标签推荐。把 TN 主题模型应用到其他类型的数据如博客和出版数据中也是很有趣的。

致谢

我们要感谢匿名审稿人的有用的反馈意见和意见。

NICTA 是由澳大利亚政府通过通信部和澳大利亚研究理事会通过卓越计划 ICT 中心成立的。

参考文献

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., and Wang, L. (2013). How noisy social media text, how diffrnt social media sources? *IJCNLP*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *the Journal of machine Learning research*, 3:993–1022.

Buntine, W., Du, L., and Nurmi, P. (2010). Bayesian networks on dirichlet distributed vectors. pages 33–40.

Chang, J. and Blei, D. M. (2010). Hierarchical relational models for document networks. *The Annals of Applied Statistics*, 4(1):124–150.

Chen, C., Du, L., and Buntine, W. (2011). Sampling table configurations for the hierarchical Poisson-Dirichlet process. In *Machine Learning and Knowledge Discovery in Databases*, pages 296–311. Springer.

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia. Association for Computational Linguistics.

Goldwater, S., Griffiths, T., and Johnson, M. (2006). Interpolating between types and tokens by estimating power-law generators. *Advances in neural information processing systems*, 18:459.

Johnson, M., Griffiths, T. L., and Goldwater, S. (2007). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. *Advances in neural information processing systems*, 19:641.

Liu, Y., Niculescu-Mizil, A., and Gryc, W. (2009). Topic-link LDA: joint models of topic and author community. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 665–672. ACM.

Lloyd, J., Orbanz, P., Ghahramani, Z., and Roy, D. (2012). Random function priors for exchangeable arrays with applications to graphs and relational data. In *Advances in Neural Information Processing Systems 25*, pages 1007–1015.

Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30. Association for Computational Linguistics.

Mehrotra, R., Sanner, S., Buntine, W., and Xie, L. (2013). Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *The 36th Annual ACM SIGIR Conference*, page 4, Dublin/Ireland.

Miller, K., Jordan, M. I., and Griffiths, T. L. (2009). Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems*, pages 1276–1284.

Nallapati, R., Ahmed, A., and Xing, E. P. (2008). Joint latent topic models for text and citations. In KDD.

Qiu, M., Zhu, F., and Jiang, J. (2013). It is not just what we say, but how we say them: Lda-based behavior-topic model. 2013 SIAM International Conference on Data Mining (SDM'13).

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494.

AUAI Press.

Sato, I. and Nakagawa, H. (2010). Topic models with power-law using pitman-yor process. In *Proceedings ofthe 16th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '10*, pages 673–682, New York, NY, USA. ACM.

Teh, Y. W. and Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press.

Tsai, F. S. (2011). A tag-topic model for blog mining. *Expert Systems with Applications*, 38(5):5330–5335.

Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web, WWW '13*, pages 1445–1456, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Yang, J. and Leskovec, J. (2011). Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European conference on Advances in information retrieval, ECIR'11*, pages 338–349, Berlin, Heidelberg. Springer-Verlag.