Computer Science and Application 计算机科学与应用, 2017, 7(3), 232-239
Published Online March 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.73029

Hans汉斯

Relation Topic Model Based on Links

Quanmin Wang, Yanfeng Sun, Zhenguo Li, Shi Gu, Kaiyang Wang

Faculty of Information Technology, Beijing University of Technology, BJUT, Beijing
Email: sunyanfeng0913@163.com

Received: Mar. 7th, 2017; accepted: Mar. 25th, 2017; published: Mar. 28th, 2017

Abstract

LDA model is a complete model of probability topic generation, which can use effective probability algorithm to train and use model. However, LDA model does not consider the effect of link between documents on topic generation in training process. In the RTM model of this paper, the link between the documents is added to the calculation, and the calculation process uses the EM algorithm to calculate the potential variables. Because it cannot be accurately calculated, the variational distribution algorithm is used. Finally, we predict the data without training, according to the document prediction link and the link prediction document. We can see that the two RTM models in the dataset are all good.

Keywords

LDA Model, RTM Model, Links, Variational Distribution

基于链接的关系主题模型

王全民，孙艳峰，李振国，谷　实，王开阳

北京工业大学信息学部计算机学院，北京
Email: sunyanfeng0913@163.com

收稿日期：2017年3月7日；录用日期：2017年3月25日；发布日期：2017年3月28日

摘　要

LDA模型是一种完全的概率主题生成模型，可以利用有效的概率算法来训练和使用模型，但是LDA模型在训练过程中并没有考虑文档之间的链接对主题生成的影响，而在本文提出的RTM模型中，就加入对文档之间链接的计算，计算过程中使用EM算法来对潜在变量进行计算，因为无法准确计算，所以采用变分分布算法。最后我们对没有训练的数据进行预测，分别为根据文档预测链接和根据链接预测文档，在试验结果中可以看到在数据集中两个RTM模型的变型都表现良好。

文章引用: 王全民, 孙艳峰, 李振国, 谷实, 王开阳. 基于链接的关系主题模型[J]. 计算机科学与应用, 2017, 7(3): 232-239. https://doi.org/10.12677/csa.2017.73029

关键词

LDA模型，RTM模型，链接，变分分布

1. 背景介绍

近年来，关于文本过滤的研究都集中在过滤模型方面，基本上是引入和改进机器学习领域的相关成果[1]。隐含狄利克雷分布[2]是近年来这方面发展起来的一种重要的离散数据集合的建模方法，首先，LDA模型是完全的概率生成模型[3]，因此具有丰富的内在结构，并且可以利用成熟有效的概率算法来训练和使用模型，再者，LDA 模型参数空间的规模是 $K*N$ (k 是隐含主题的数量，N 是此表中词的数量)，与文档的数量无关，使得 LDA 更适合在大规模语料库上构造文本表示模型。但是 LDA 模型只考虑文本本身的内容，而忽略了节点之间的联系。

2. 相关研究

这个关系主体模型(RTM)是在统计和机器学习的研究基础上提出的，它是一个可以提供节点属性以及其网络结构的潜在空间模型，一些对网络结构的潜在空间模型已经被提出过[4]，但是，这些模型只单单计算数据，并没有考虑节点属性；解释网络的链接结构或者是节点属性的模型的建立过程大都是通过降维[5]和获取节点属性[6]来建立，都趋向于研究它们中的一个或者它们之外的其他属性[7]，包含一些主题模型[8]，RTM 模型综合考虑了节点属性和他们之间的链接结构，这样可以通过其中一个预测到另一个。

对一篇文档来说，每一个节点信息就是它包含的关键词，RTM 模型研究的是文档关键词和文档之间的链接的关系。除了可以通过链接预测词语和词语预测链接，还可以对没有经过训练的文档数据进行预测。RTM 模型是一种新的针对文档和文档之间链接的概率生成模型[9]，关于生成模型[10]，在之前研究中，通常把链接当作相互独立的单元，与本文提出的研究最相近的是 Nallapati 和 Mei 的研究[11]，他们试图扩展混合成员模型[12]，他们假设可交换，在模型中允许使用主题来解释链接并且使用其他词语来解释另外一些词语，但是这样影响了使用词语信息来预测链接信息，与此不同的是，在 RTM 模型中强制使用主题来解释全部词语和链接。RTM 是一种新的概率生成模型，它可以被用来分析链接集，比如网页引用、网页链接、社交网络等，我们在实验中证实了它在分析这些数据的时候是适用的，与之前的模型相比在效果上有了明显的提升。

3. RTM 模型

3.1. RTM 简介

RTM 假设一组被观察到的文档 $w_{1:D,1:N}$ 以及他们之间的双链接 $y_{1:D,1:D}$ 是由以下方法产生的，见图 1。

1.对于每一个文档 d：

a.抽取主题比例 $\theta_d|\alpha \sim \mathrm{Dir}(\alpha)$

b.对于每一个词语 $w_{d,n}$

i.抽取主题分配 $z_{d,n}|\theta_d \sim \mathrm{Mult}(\theta_d)$

ii.抽取单词 $w_{d,n}|z_{d,n},\beta_{1:K} \sim \mathrm{Mult}(\beta z_{d,n})$

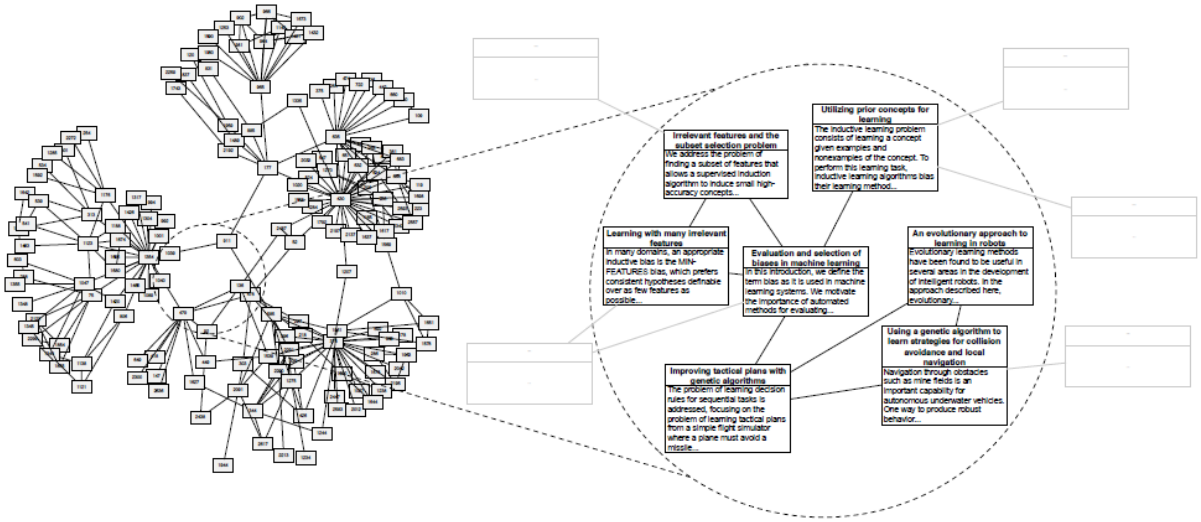

Figure 1. Graphical model of a large number of files
图 1. 大量文件的图形模型

2.对于每一对文档 d,d'：

a.抽取二元连接指示器 $y|z_d,z_{d'} \sim \psi\left(.|z_d,z_{d'}\right)$

完整的模式是很难说明的。因为其中包含了所有从文档中观察的词语，以及他们之间的每个可能的链接变量。函数 $\psi\left(.|z_d,z_{d'}\right)$ 是两个文件之间的链接分布。这个函数依赖于生成他们的词语 $z_d,z_{d'}$，这里我们探讨两种可能性。

第一，我们考虑：

$$\psi_\sigma\left(y=1\right)=\sigma\left(\eta^{\mathrm{T}}\left(z_d\circ z_{d'}\right)+\nu\right) \tag{1}$$

其中 $z_d=\frac{1}{N_d}\sum_n z_{d,n}$，$\sigma$ 表示函数是 S 状的曲线。这个链接函数对每一对二元变量作隐藏协变量回归建模。它由 η 参数化、由 υ 拦截。协变量是由 $z_d\circ z_{d'}$ 构造，捕捉两个文件隐含主题之间的相似性。

第二：我们考虑：

$$\psi_e\left(y=1\right)=\exp\left(\eta^{\mathrm{T}}\left(z_d\circ z_{d'}\right)+\upsilon\right) \tag{2}$$

ψ_e 使用和 ψ_σ 相同的协变量，但是由一个指数函数代替，这个函数返回的概率呈指数增长，可以被看作是 Blei 提出的建模方法[13]的一个近似变体。

下面我们对这两种情况进行分析。

3.2 计算过程

图 2 表示了文档间的关系。

变量 y 表示两个文件是否有联系。

通过上面的 ψ 函数可以认为，潜在特性的期望函数是回应 $z_d\circ z_{d'}$ 这个公式，由监督 LDA 模型[14]可确保用于生成文档内容的相同潜在主题分配，另外还负责生成他们的链接结构。

3.2.1. 算法建模

我们的计算过程参考变分[15]推理过程，在变分推理中使用 EM (期望最大化)算法来做参数估计，最大期望(EM)算法是在概率模型[16]中寻找参数极大似然估计或者最大后验估计的算法，其中概率模型依赖于无法观测到的潜在变量，那么就是计算观测变量的潜在变量的后验分布，但是准确的后验分布是难

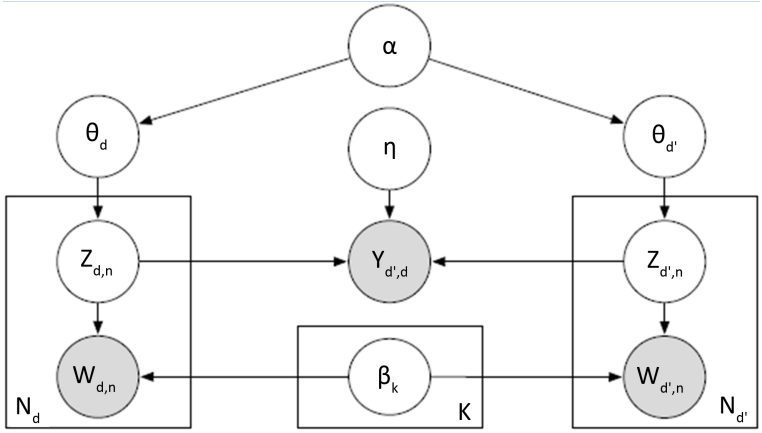

Figure 2. Document relationship diagram
图 2. 文档关系图

以计算的，所以只能使用变分分布。在变分分布中，首先假设存在一个由参数索引的潜在变量的分布函数，这些参数接近真实的后验，然后使用相对熵测量(见 Jordan 等，1999) [17]。经过分解后的 γ 是一组狄利克雷参数，每一个对应一个文档；而 Φ 是一组多项参数，对应每个文档中每个词语。

$$q\left(\Theta,\mathrm{Z}|\gamma,\Phi\right)=\prod_d\left[q_\theta\left(\theta_d|\gamma_d\right)\right]\prod_n q_z\left(z_{d,n}|\phi_{d,n}\right) \tag{3}$$

这个推理过程只对观测到的链接建模，即 $y_{d_1,d_2}=1$，这样做的原因有两个：

首先，当文档 $d1$ 和 $d2$ 之间的链接被观测到就修改 $y_{d_1,d_2}=1$，否则 $y_{d_1,d_2}=0$，但是这种方法不能证明当 $y_{d_1,d_2}=0$ 时，$d1$ 和 $d2$ 之间没有链接，此时，对待这些链接作为潜在变量是更符合真实的，例如，在大型的社交网络如 Facebook 中，两个人之间没有联系并不一定意味着他们不是朋友，他们可能是真正的朋友，只是谁都不知道在彼此的存在而以[18]。

第二，隐藏未被观测到的链接可以降低计算成本，因为计算的复杂性与观察到的链接的数量成正比。

我们选择等式一时，由于等式难以计算，对它进行一次近似[19]，得到等式 4 于是乎目标转化为计算等式 4。

$$\begin{aligned}\zeta=&\sum_{(d_1,d_2)}E_q\left[\log p\left(y_1,y_2|z_{d_1},z_{d_2},\eta,\upsilon\right)\right]\\&+\sum_d\sum_n E_q\left[\log p\left(w_{d,n}|\beta_{1:K},z_{d,n}\right)\right]\\&+\sum_d\sum_n E_q\left[\log p\left(z_{d,n}|\theta_d\right)\right]\\&+\sum_d E_q\left[\log p\left(\theta_d|\alpha\right)\right]+H\left(q\right)\end{aligned} \tag{4}$$

$$\begin{aligned}\zeta_{d_1,d_2}&\equiv E_q\left[\log p\left(y_1,y_2=1|z_{d_1},z_{d_2},\eta,\upsilon\right)\right]\\&\approx\eta^{\mathrm{T}}\bar{\pi}_{d_1,d_2}+\upsilon+\log\sigma\left(-\eta^{\mathrm{T}}\bar{\pi}_{d_1,d_2}-\upsilon\right)\end{aligned} \tag{5}$$

其中 $\bar{\pi}_{d_1,d_2}=\overline{\phi_{d_1}}\circ\overline{\phi_{d_2}}$ 以及 $\bar{\phi}_d=E_q\left[\overline{\mathrm{Z}_d}\right]=\frac{1}{N_d}\sum_n\phi_{d,n}$，当 ψ_e 是目标函数时，这个式子可以被准确的计算为：

$$E_q\left[\log p\left(y_{d_1,d_2}=1|\bar{z}_{d_1},\bar{z}_{d_2},\eta,\upsilon\right)\right]=\eta^{\mathrm{T}}\bar{\pi}_{d_1,d_2}+\upsilon \tag{6}$$

使用坐标上升方法来优化变分参数 γ 和 ϕ 可得到：

$$\phi_{d,j}\propto\exp\left\{\sum_{d'\neq d}\left(\nabla_{\pi_{d,d'}}\zeta_{d,d'}\right)\frac{\eta\circ\overline{\phi_{d'}}}{N_d}+E_q\left[\log\theta_d|\gamma_d\right]+\log\beta_{\cdot,w_{d,j}}\right\}$$

$\zeta_{d,d'}$ 的计算取决于方程式 5 或者 6 中对于 ψ 的选择。$\log \beta_{\cdot,w_{d,j}}$ 可以通过对 $w_{d,j}$ 元素取对数被计算。$E_q\left[\log \theta_d \middle| \gamma_d\right]$ 是 $\Psi(\gamma_d)-\Psi\left(\sum \gamma_{d,i}\right)$，而 Ψ 是双伽马函数。

γ 的更新和 LDA 模型中变分参数一样。即 $\gamma_d \leftarrow \alpha+\sum_n \phi_{d,n}$。

3.2.2. 估算参数

我们通过对每个参数计算其极大似然估计的方法来对模型进行调整，主要是多项主题向量 $\beta_{1:K}$ 和链接函数参数 η、υ，但是我们发现直接计算是比较困难的，所以我们转向求近似值，采用变分 EM，在优化式子 4 的变分分布和模型参数之间迭代。

因为包含 β 的式子 4 中术语和传统 LDA 中的一样，所以估算主题向量可以通过相同的方法，即对称的狄利克雷来估算 $\beta_{k,w}$。

$$\beta_{k,w} \propto \sum_d \sum_n \left(w_{d,n}=w\right)\phi_{d,n}^k$$

但是不能在没有经过观察的情况下直接优化链路概率函数的参数，而是通过使用分等级的正则化任意参数化，这种正则化是先假设网络中存在一些潜在的负面意见，然后将其纳入参数估计，负观测的频率由 ρ 控制。当使用公式 1 的逻辑，我们使用基于梯度的优化[20]来估计参数 η 和 υ。使用公式 5 中使用的近似，ELBO 的相关梯度是

$$\triangleright_\eta\, \zeta \approx \sum_{(d_1,d_2)}\left[1-\sigma\left(\eta^{\mathrm{T}}\overline{\pi}_{d_1,d_2}+\upsilon\right)\right]\overline{\pi}_{d_1,d_2}-\rho\sigma\left(\eta/K^2+\upsilon\right)/K^2,$$

$$\frac{\partial}{\partial \upsilon}\zeta \approx \sum_{(d_1,d_2)}\left[1-\sigma\left(\eta^{\mathrm{T}}\overline{\pi}_{d_1,d_2}+\upsilon\right)\right]-\rho\sigma\left(1^{\mathrm{T}}\eta/K^2+\upsilon\right).$$

当使用等式 2 的指数函数时，可以分析出参数 η 和 υ。

$$\upsilon \leftarrow \log\left(1-1^{\mathrm{T}}\overline{\Pi}\right)-\log\left(\rho\frac{K-1}{K}+1-1^{\mathrm{T}}\overline{\Pi}\right)$$

$$\eta \leftarrow \log\left(\overline{\Pi}\right)-\log\left(\overline{\Pi}+\frac{\rho}{K^2}1\right)-1\upsilon,$$

其中，$\overline{\Pi}=\sum_{(d_1,d_2)}\overline{\pi}_{d_1,d_2}$。

我们的最终目标是要对新的数据进行预测，分为两个类别的预测：从文字预测链接的和从链接预测文字。

在链接预测中，给定一个新的文档及其中的词语。我们需要从这个文档到其他文件的链接。这需要计算一个关于后验的期望，这个后验我们是无法计算的。

$$p\left(y_{d,d'}\middle|w_d,w_{d'}\right)=\sum_{z_d,z_{d'}} p\left(y_{d,d'}\middle|\overline{z}_d,\overline{z}_{d'}\right)p\left(z_d,z_{d'}\middle|w_d,w_{d'}\right)$$

通过之前介绍的推理算法，根据变分分布的优化方法，我们得到了使用训练文本集[21]中的词语和链接以及测试文件中的词语来进行计算的方法，使用近似值 $q(\Theta,Z)$ 来替换上面提到的后验，那么预测是值就约等于：

$$p\left(y_{d,d'}\middle|w_d,w_{d'}\right)\approx E_q\left[p\left(y_{d,d'}\middle|\overline{z}_d,\overline{z}_{d'}\right)\right] \tag{7}$$

在词语预测中，我们仅仅基于链接预测一篇未知文档中的词语，与链路预测一样，$p\left(w_{d,i}\middle|y_d\right)$ 无法计算，使用和上面相同的技术，使用变分分布来近似这个后验，就产生了预测等式：

$$p\left(w_{d,i}\middle|y_d\right)\approx E_q\left[p\left(w_{d,i}\middle|z_{d,i}\right)\right]$$

通过文件和链接，我们的模型能够根据词语预测链接，以及根据链接预测词语，或者两者混合。

4. 实验结果

我们使用一组数据集来做实验，完成分词处理的词语(删除停顿词和常用词)，由定向链接转化为不定向链接，删除没有直接链接的文件[22]，Cora 数据[23]包含使用论文搜索引擎搜索出来的摘要，以及文件间相互引用的链接[2]。

评估

RTM 模型对未经训练的数据定义了概率分布，根据第三部分所描述的，从数据中推断潜在变量，我们需要知道的是在预测未经训练的数据时这个模型的效果有多好，我们研究上面所说的 RTM 的两种变体：

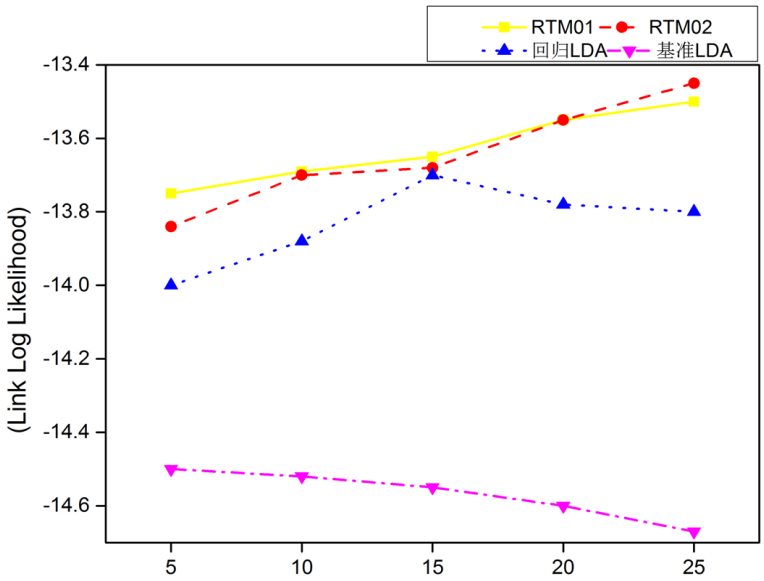

Figure 3. Link prediction results comparison chart
图 3. 链接预测结果对比图

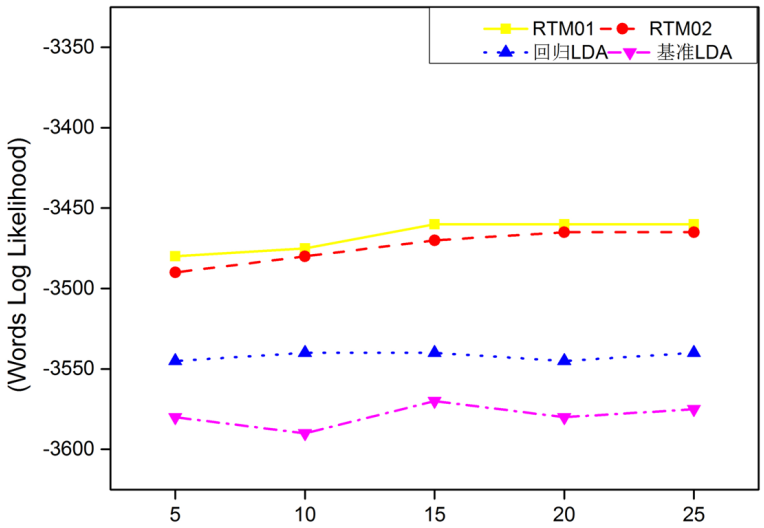

Figure 4. Word prediction results comparison chart
图 4. 词语预测结果对比图

利用等式 1 使用逻辑链接的逻辑 RTM01 和利用等式二使用指数链接的指数 RTM02，通过两种备选方案来比较这两种模型，首先是基准模型，在这个模型中词语和链接是相互独立的，使用多项分布对词语建模，使用伯努利对链接进行建模；第二种是回归 LDA，首次拟合 LDA 模型对文件处理，然后对观测到的链接进行逻辑回归，并输入每对文件的潜在链接，而不是进行降维和回归，该方法首先进行无监督降维，然后回归潜在空间和潜在的连结结构之间的关系。所有的模型都进行了训练，α 的值取 5.0。

通过对上面所说模型关于词语预测和链接预测的计算，我们将数据集分成四份，对于每一份数据对应一种模型，我们提出两个预测查询：给定一个未知文档中的词语，他们之间的链接是怎样的；给定文档中的链接，它包含的词语有哪些。其次，预测查询是针对未经过观测训练的文件，在训练测试文档时，它们之间的链接，在图 3 中显示了链接预测的结果，图 4 显示的是词语预测的结果。

在预测链接上，RTM 的两个变体比另外两个模型在所有数据集上表现都要好，Cora 数据集是例证，指数 RTM 比基准模型提高了 6%，比回归 LDA 提高了 5%；逻辑 RTM 比基准模型提高了将近 5%，比回归 LDA 提高了 4%。在词语预测上，RTM 的两个变体又一次的比其他模型都要好，这是因为 RTM 模型使用链接信息去影响词语的预测分布，LDA 回归模型和预测和基准模型相似。

参考文献 (References)

[1] 李文波, 孙乐, 黄瑞红. 基于 Labeled-LDA 模型的文本分类新算法[J]. 计算机学报, 2008, 31(4): 620-627.

[2] 张启蕊, 张凌, 董守斌, 谭景华. 训练集类别分布对文本分类的影响[J]. 清华大学学报(自然科学版), 2005, 45(S1): 76-79.

[3] Blei, D., Ng, A. and Jordan, M. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993-1022.

[4] Cohn, D. and Hofmann, T. (2011) The Missing Link—A Probabilistic Model of Document Content and Hypertext Connectivity. *Neural Information Processing Systems*.

[5] Newman, M. (2012) The Structure and Function of Networks. *Computer Physics Communications*, **147**, 40-45.

[6] Taskar, B., Wong, M., Abbeel, P. and Koller, D. (2014) Link Prediction in Relational Data. *Neural Information Processing Systems*.

[7] McCallum, A., Nigam, K., Rennie, J. and Seymore, K. (2009) Automating the Construction of Internet Portals with Machine Learning. *Information Retrieval*, **3**, 127-163. https://doi.org/10.1023/A:1009953814988

[8] McCallum, A., Corrada-Emmanuel, A. and Wang, X. (2010) Topic and Role Discovery in Social Networks. *IJCAI*.

[9] Nallapati, R., Ahmed, A., Xing, E.P. and Cohen, W.W. (2008) Joint Latent Topic Models for Text and Citations. *Proceedings of the* 14*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, 24-27 August 2008, 542-550. https://doi.org/10.1145/1401890.1401957

[10] Nallapati, R. and Cohen, W. (2013) Link-pLSA-LDA: A New Unsupervised Model for Topics and Influence of Blogs. 2*nd International Conference on Weblogs and Social Media* (*ICWSM*), Seattle, 2008.

[11] Mei, Q., Cai, D., Zhang, D. and Zhai, C. (2008) Topic Modeling with Network Regularization. *Proceedings of the 17th international conference on World Wide Web*, Beijing, 21-25 April 2008, 101-110. https://doi.org/10.1145/1367497.1367512

[12] Airoldi, E., Blei, D., Fienberg, S. and Xing, E. (2008) Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, **9**, 1981-2014.

[13] Blei, D. and Jordan, M. (2013) Modeling Annotated Data. *Proceedings of the* 26*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, 28 July-1 August 2003, 127-134.

[14] Blei, D.M. and McAuliffe, J.D. (2007) Supervised Topic Models. *Neural Information Processing Systems*.

[15] Braun, M. and McAuliffe, J. (2010) Variational Inference for Large-Scale Models of Discrete Choice. *Journal of the American Statistical Association*, **105**, 324-334. arXiv:0712.2526. https://doi.org/10.1198/jasa.2009.tm08030

[16] Sinkkonen, J., Aukia, J. and Kaski, S. (2008) Component Models for Large Networks. arXiv:0803.1628

[17] Gruber, A., Rosen-Zvi, M. and Weiss, Y. (2008) Latent Topic Models for Hypertext. *Proceedings of the* 24*th Conference on Uncertainty in Artificial Intelligence*.

[18] Hoff, P., Raftery, A. and Handcock, M. (2012) Latent Space Approaches to Social Network Analysis. *Journal of the*

American Statistical Association, **97**, 1090-1098. https://doi.org/10.1198/016214502388618906

[19] Getoor, L., Friedman, N., Koller, D. and Taskar, B. (2011) Learning Probabilistic Models of Relational Structure. *Proceedings of the* 18*th International Conference on Machine Learning*, 28 June-1 July 2001, 170-177.

[20] Wang, X., Mohanty, N. and McCallum, A. (2005) Group and Topic Discovery from Relations and Text. *Proceedings of the* 3*rd International Workshop on Link Discovery*, Chicago, 21-25 August 2005, 28-35. https://doi.org/10.1145/1134271.1134276

[21] Hofman, J. and Wiggins, C. (2008) Bayesian Approach to Network Modularity. *Physical Review Letters*, **100**, Article ID: 258701. arXiv:0709.3512. https://doi.org/10.1103/physrevlett.100.258701

[22] Dietz, L., Bickel, S. and Scheffer, T. (2007) Unsupervised Prediction of Citation Influences. *Proceedings of the* 24*th International Conference on Machine Learning*, Corvalis, Oregon, 20-24 June 2007, 233-240. https://doi.org/10.1145/1273496.1273526

[23] Kemp, C., Griffiths, T. and Tenenbaum, J. (2014) Discovering Latent Classes in Relational Data. MIT AI Memo 2004-019.