Modeling Structured Interaction in Large Online User Forums

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present here an approach to model online social forums that respects the structure of the discussion and thus inturn provides researcher with unique insights into the myriads of user forums in the todays online communties. We bring together the structure of the forum network as well as texts posted in a model that respects the structure of the conversation in the forum. The work also focuses on large scal interaction and provides an efficient approximate estimation technique that is scalable i am thinking of using stochastic variational for scalability if time permits . Analysing Wikipidia edit and cancer patient online user forums using this technique provides us interesting insights into these online user communities.

1 Introduction

There have been flood of online forums in recent decade and consequently so have been the focus of academic research and industry on online social networks. Analysing online social networks and user forums have been approached using various perspective such as graph/network based [11, 10] , probabilistic graphical model [1], combined network & text mining based [7, 8] approaches. But very few of the approaches have taken into the structural framework in which the conversation in online forums happen. This is important to correctly model the interaction as well as the content posted by the users while interacting with the user community. E.g. in an onlne forum there are topic-threads and users post their responses on this thread after possibly reading trhough the responses of other users in this thread. And the users possibly posts multiple times on the thread in froms of replies to other posts in the thread. For analysing the user interaction, besides taking into account the user interaction network and the text posted, it becomes imperative that the structure of the conversation must also be taken into account. This will enable us to gain deeper insights into user behavior in the online community. Very few research work has tried to bring the forum structure in the analysis of online communities.Our work here besides bringing network modeling and text mining together puts in the forum structure in the model to provide a more robust analysis of the user interactions.

The current work also focuses on analysing large scale user interaction in big online social forums. We provide a variational approximation based estimation technique that is scalable to big forums with thousands of users. Also write about stochastic approximation if given we have time and we can get it working

2 Related Work

Our project lies at the intersection of graph clustering and social emergence. Wikipedia's talk pages are an instance of a large social community where we can observe social emergence. Sociologists

divide emergents into various levels [9] e.g. Individual-specific interactions, Ephemeral emergent interactions, Stable emergent interaction etc.

For role-identification and clustering users based on roles in online communities, White et al.[13] proposed a mixed-membership model that obtained membership probabilities for discussion-forum users for each statistic (in- and out-degrees, initiation rate and reciprocity) in various profiles and clustered the users into "extreme profiles". Ho et al. [7] presented TopicBlock that combines text and network data for building a taxonomy for a corpus.

Griffiths et al. [5] described a generative model for Blei et al.'s LDA model [2] using an MCMC algorithm for bayesian inference queries on the model.

The LDA model and MMSB models were combined by Nallapati et al. [8] using the Pairwise-Link-LDA and Link-PLSA-LDA models where documents are assigned membership probabilities into bins obtained by topic-models.

For simultaneously modeling topics in bilingual-corpora, Smet et al. [12] proposed the Bi-LDA model which generates topics from the target languages for paired documents in these very languages. The end-goal of their approach is to classify any document to one of the obtained set of topics.

For modeling the behavioral aspects of entities and discovering communities in social networks, several game-theoretic approaches have been proposed (Chen et al. [3], Yadati and Narayanam [14]).

Our work is unique in this context as it tries to bridge the gap between community discovery and social emergence. A very popular approach for network clustering is to use a generative framework to infer the underlying structure of communities in a graph. Airoldi et al.[1] proposed a mixed membership stochastic block model to infer community structure in a network. Here they use a generative scheme which captures membership of a user to multiple communities using latent variables. There are other variants of mixed membership stochastic models, e.g. Ho et al.[6] that study evolving clusters over time varying networks while Fu et al.[4] explore dynamic mixed membership models.

3 Approach

Online forums have a generally a specific structure that provides a lot of context to all the interactions among the users. Ignoring this in the analysis makes us lose a lot of precious information as we will see in later sections. Here we describe a typeical forum and the answers that we plan to obtain.

3.1 structure in online forums

We discuss here the ideas and questions that are important for modeling structure in online forums. Following are the questions that we want answered:

1. When two persons interact in a thread or a post which topic/community they each belong to
2. When a user U is in communnity C what type of text does he use to communicate
3. Multi-user-Interaction: In a thread a user can post by addressing to a specific user but he is also talking to other users in the thread simultaneously. Can we model this phenomenon
4. There is an inherent bias towards the thread starter or in turn topic of the theread; can such an information be utilised in some form of a prior value/input
5. Multi-layer-Interaction: On the network side of things there are multiple signals which cannt be simply added to make a single signal e.g. different types of edges in the graphs (user calling by username and nick-name). Can the model take this into account. We are not doing this at present.
6. User posts aggregation; there are multiple ways to aggreagte
   (a) Network Layer aggregation: We call all types of edges as a single edge type and use this combined signal.
   (b) aggregating user posts accross multiple threads in the forum.

(c) Aggregating user post only in the same thread

(d) Aggregating user post only for same user-user pair interaction; i.e. a user might have posted multiple replies to another user and we aggregate all such replies into one for this user pair interaction.

(e) No aggregation at all.

3.2 graphical model & generative story

Based on the discussions above we came up with the following final model shown in figure 1. In this model, figure 1 below, we aggregate the posts of a given user in a given thread into one document called R_p.

The generative process for the figure is as follows:

Assuming that there are total N_t users in the thread t.

- For each Thread t
  - For each user $p \in \mathcal{N}_t$
    - Draw a K dimensional mixed membership vector $\vec{\pi}_p \sim$ Dirichlet(α)
    - Draw $B(g,h) \sim Gamma(\kappa, \eta)$; where κ, η are parameters of the gamma distribution.
  - For each pair of users $(p,q) \in \mathcal{N}_t \times \mathcal{N}_t$:
    - Draw membership indicator for the indicator, $\vec{z}_{(p\to q,t)} \sim$ Multinomial(π_p).
    - Draw membership indicator for the receiver, $\vec{z}_{(q\to p,t)} \sim$ Multinomial(π_q).
    - Sample the value of their interaction, $Y(p,q,t) \sim$ Poisson($\vec{z}^{\top}_{(p\to q,t)} B \vec{z}_{(p\leftarrow q,t)}$).
  - For each user $p \in \mathcal{N}_t$
    - Draw ϕ_k from $Dirichlet(\beta)$.
    - Form the set $Q_{p,t}$ that contains all the users that p interacts to on thread t
      - For each word $w \in R_{p,t}$
      - Draw $w \sim \phi(w|z_{(p\to q,t)}, \forall q \in Q_{p,t})$

The data likelihood for the model in figure 1

$$P(Y, R_p|\alpha,\beta,\kappa,\eta) = \int_\Phi \int_\Pi \sum_z P(Y, R_p, z_{p\to q}, z_{p\leftarrow q}, \Phi, \Pi|\alpha,\beta,\kappa,\eta)$$

$$= \int_\Phi \int_\Pi \sum_z \Big[\prod_{p,q} \prod_t P(Y^t_{pq}|z^t_{p\to q}, z^t_{p\leftarrow q}, B) \cdot P(z^t_{p\to q}|\Pi_p) \cdot P(z^t_{p\leftarrow q}|\Pi_q)$$

$$\cdot \left(\prod_p P(\Pi_p|\alpha) \prod_t \prod_p P(R^t_p|z^t_{p\to q}, \Phi) \cdot \prod_k P(\Phi_k|\beta) \right) \cdot \prod_{g,h} P(B_{gh}|\eta,\kappa) \Big] \tag{1}$$

The complete log likeliood of the model is:

$$\log P(Y, W, z_\to, z_\leftarrow, \Phi, \Pi, B|\kappa,\eta,\beta,\alpha) = \sum_t \sum_{p,q} \log P(Y^t_{pq}|z^t_{p\to q}, z^t_{p\leftarrow q}, B) +$$

$$\sum_t \sum_{p,q} (\log P(z^t_{p\to q}|\Pi_p) + \log P(z^t_{p\leftarrow q}|\Pi_p)) + \sum_p \log P(\Pi_p|\alpha) +$$

$$\sum_t \sum_p \sum_{w\in R^t_p} \log P(w|z_{p\to}, \Phi) + \sum_k \log P(\Phi_k|\beta) + \sum_{gh} \log P(B_{gh}|\eta,\kappa) \tag{2}$$

The mean field variational approximation for the posterior is

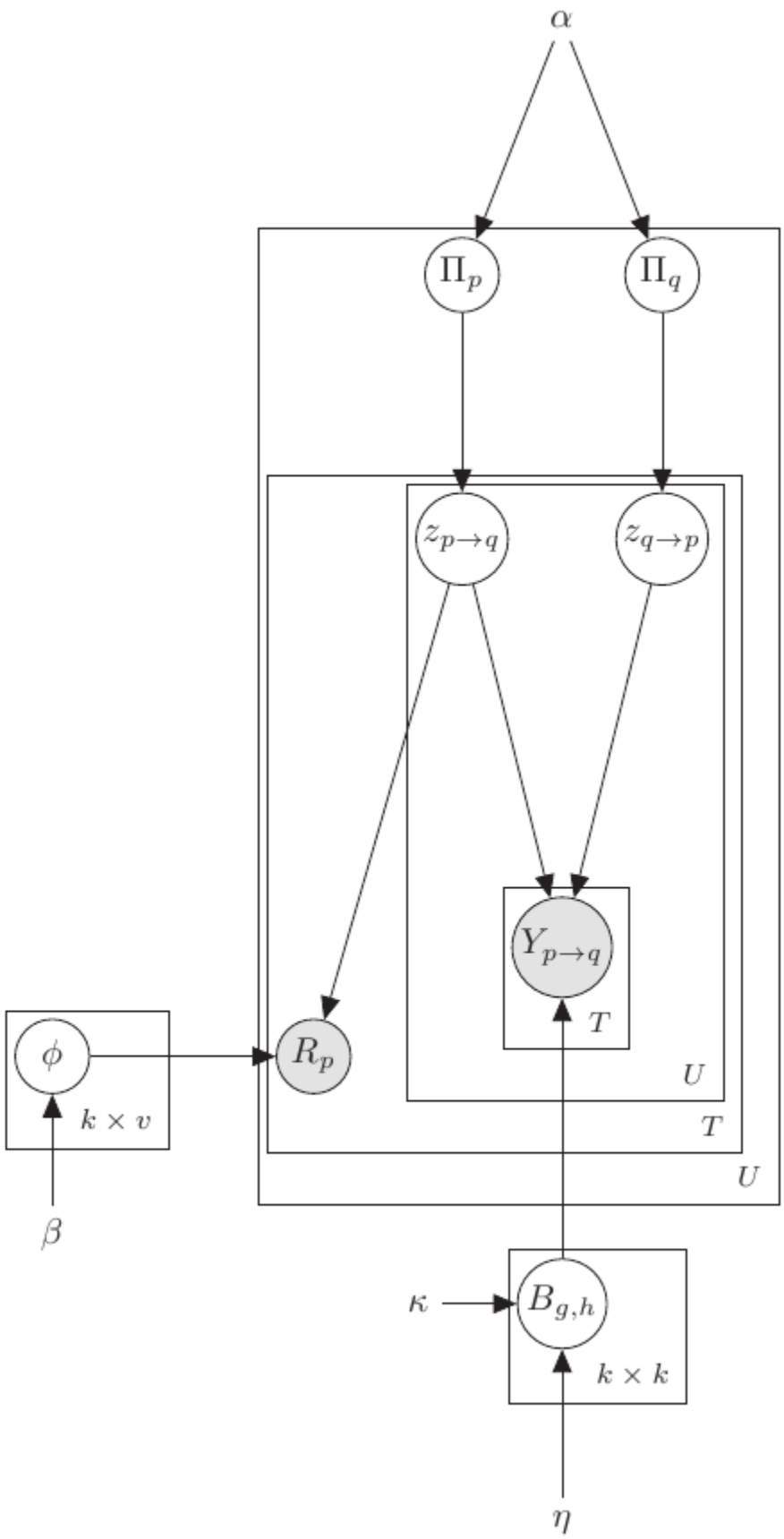

Figure 1: This graphical model takes into account multi-way interaction among users in a thread simultaneously

$$q(z,\Phi,\Pi,B|\Delta_{z_\rightarrow},\Delta_\Phi,\Delta_B,\Delta_{z_\leftarrow},\Delta_{B_\kappa}) = \prod_t\prod_{p,q}\Big(q_1(z^t_{p\rightarrow q}|\Delta_{z_{p\rightarrow q}}) + q_1(z^t_{p\leftarrow q}|\Delta_{z_{p\leftarrow q}})\Big) \cdot \prod_p q_4(\Pi_p|\Delta_{\Pi_p})\prod_k q_3(\Phi_k|\Delta_{\Phi_k})\prod_{g,h} q(B_{g,h}|\Delta_{B_\eta},\Delta_{B_\kappa}) \quad (3)$$

The lower bound for the data log-likelihood from jensen's inequality is:

$$L_\Delta = E_q\Big[\log P(Y,W,z_\rightarrow,z_\leftarrow,\Phi,\Pi,B|\kappa,\eta,\beta,\alpha) - \log q\Big] \quad (4)$$

$$\begin{aligned} L_\Delta = E_q\Bigg[&\sum_t\sum_{p,q}\log\Big(B_{g,h}^{Y^t_{p,q}}\frac{e^{-B_{gh}}}{Y^t_{pq}!}\Big) + \sum_t\sum_{pq}\log\Big(\prod_k(\pi_{p,k}^{z_{p\rightarrow q}=k})\Big) + \sum_t\sum_{p,q}\log\Big(\prod_k(\pi_{q,k})^{z_{p\leftarrow q}=k}\Big) + \\ &\sum_p\log\Big[\prod_k(\Pi_{p,k})^{\alpha_k-1}\cdot\frac{\Gamma(\sum\alpha_k)}{\prod_k\Gamma(\alpha_k)}\Big] + \sum_t\sum_p\sum_{w\in R^t_p}\log\Big(\prod_{u\in V}(\bar{z}^T\phi_u)^{w=u}\Big) + \\ &\sum_k\log\Big[\prod_{u\in V}(\phi_{k,u})^{\beta_k-1}\cdot\frac{\Gamma(\sum\beta_k)}{\prod_k\Gamma(\beta_k)}\Big] + \sum_{g,h}\log\Big(B_{g,h}^{\kappa-1}/\eta^\kappa\Gamma(\kappa)\cdot\exp(-B_{g,h}/\eta)\Big) \\ &- E_q\Bigg[\sum_t\sum_{p,q}\log\big(\prod_k(\Delta_{z_{p\rightarrow q},k})^{z_{p\rightarrow q}=k}\big) + \sum_t\sum_{p,q}\log\big(\prod_k(\Delta_{z_{p\leftarrow q},k})^{z_{p\leftarrow q}=k}\big) + \\ &\sum\log\big[\prod_k(\Pi_{p,k})^{\Delta_{\pi_{pk}}-1}\frac{\Gamma(\Delta_{\Pi_p})}{\prod_{k=1}\Gamma(\Delta_{\Pi_{p,k}})}\big] + \sum_k\log\big[\prod_{u\in v}(\Phi_{k,u})^{\Delta_{\Phi_{ku}}-1)}\frac{\Gamma(\Delta_{\Phi_k})}{\prod_{u\in v}\Gamma(\Delta_{\Phi_{k,u}})}\big] + \\ &\sum_{g,h}\log\big[\frac{B_{g,h}^{\Delta_{\kappa=1}}}{\Delta_\eta^{\Delta_\kappa}\Gamma(\Delta_\kappa)}\exp(-B_{g,h}/\Delta_\eta)\big] \end{aligned} \quad (5)$$

Equation 5 is the variational lower bound of the log likelihood function which is to be maximized. There are terms like $E_q\Big[\sum_{g,h}\log\Big(B_{g,h}^{\kappa-1}/\eta^\kappa\Gamma(\kappa)\cdot\exp(-B_{g,h}/\eta)\Big)\Big]$ which can be obtained by taking derivation of the partition function of the exponential family form of gamma distribution. I still have to figure out an effective way to evaluate $E_q\Big[\sum_t\sum_p\sum_{w\in R^t_p}\log\Big(\prod_{u\in V}(\bar{z}^T\phi_u)^{w=u}\Big)\Big]$ which Chong suggested (and Eric too in the last meeting) to evaluate by intriducing an additional latent variable $\bar{z}_p$ which is a realization of the average $\frac{\sum_{q\in Q}z_{p\rightarrow q}}{|Q|}$.

4 Datasets

5 Experiments & Results & Evaluations

6 Conclusion & Future Work

References

[1] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, June 2008.

[2] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] Wei Chen, Zhenming Liu, Xiaorui Sun, and Yajun Wang. A game-theoretic framework to identify overlapping communities in social networks. *Data Min. Knowl. Discov.*, 21(2):224–240, September 2010.

[4] Wenjie Fu, Le Song, and Eric P. Xing. Dynamic mixed membership blockmodel for evolving networks. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 329–336, New York, NY, USA, 2009. ACM.

[5] T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[6] Q. Ho, L. Song, and E. Xing. Evolving cluster mixed-membership blockmodel for time-evolving networks. In *International Conference on Artifical Intelligence and Statistics (AISTATS)*, 2011.

[7] Qirong Ho, Jacob Eisenstein, and Eric P. Xing. Document hierarchies from text and links. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 739–748, New York, NY, USA, 2012. ACM.

[8] Ramesh M. Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 542–550, New York, NY, USA, 2008. ACM.

[9] R. Keith (Robert Keith) Sawyer. *Social emergence : societies as complex systems / R. Keith Sawyer*. Cambridge ; New York : Cambridge University Press, 2005. Formerly CIP.

[10] Jianbo Shi. Learning segmentation by random walks. In *In Advances in Neural Information Processing*, pages 470–477. MIT Press, 2000.

[11] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, August 2000.

[12] Wim De Smet, Jie Tang, and Marie-Francine Moens. Knowledge transfer across multilingual corpora via latent topics. In *Proceedings of the 15th Pacific-Asia conference on Advances in knowledge discovery and data mining - Volume Part I*, PAKDD'11, pages 549–560, Berlin, Heidelberg, 2011. Springer-Verlag.

[13] Arthur White, Jeffrey Chan, Conor Hayes, and Brendan Murphy. Mixed membership models for exploring user roles in online fora, 2012.

[14] Narahari Yadati and Ramasuri Narayanam. Game theoretic models for social network analysis. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 291–292, New York, NY, USA, 2011. ACM.