Capturing Participation Profile in Large Online User Forums

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present here an approach to model large online social forums that respects the structure of the discussion and provides a participation profile for various users in present in the discussion. We bring together the structure of the forum as well as texts posted there in a model that captures the user-participation dynamics in the forum. The work also has a scalability focus and provides an efficient approximate estimation technique based off stochastic variational methods and can model very large forums with upto half a million users. Modelling cancer patients, Reddit and Stackoverflow online user discussions using this technique provides us with valuable insights into these communities. The model also performs well on link prediciotn and other prediction task further validating our claims. We should make it more strongly worded in our favor if the cancer forum task works well

1 Introduction

There have been a 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 [9, 8] , probabilistic graphical model [1], combined network & text mining based [4, 7] based approaches. But very few of these have taken into account the structural framework in which the conversation in online forums happen. This is important to correctly model the interaction as well as the contents posted by the users during their conversation 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 through the responses of other users in this thread. And the users possibly posts multiple times on the thread in the form of replies to other posts in the thread. For analysing such a user interaction it becomes imperative that the structure of the conversation must also be taken into account besides taking into account the user interaction network and the text posted. This enables us to gain deeper insights into user behavior in the online community that was not possible earlier. Very few research works have tried to bring the forum structure in the analysis of online communities. This is what set our work apart from the past works, our approach here besides bringing network modeling and text mining together adds in the forum structure in the model to provide a more robust analysis of the user interactions. The model also incorporates strength of interaction among the users by incorporating interaction counts as compared to MMSB model which just looks presence or absence of link [1]. In the process we discover interesting online communities and social phenomena.

The current work also focuses on analysing large scale user interactions in big online social forums. We provide a stochastic variational approximation [6] based estimation technique that is scalable to big forums with thousands of users.

2 Related Work

Our project lies at the intersection of community discovery, structure modelling and text mining. Wikipedia's talk pages are an instance of a large social community where we can observe users networking with each other as well as posting content in a structured way. Similar phenomena are observed across almost all social networking websites and online forums.

For role-identification and clustering users based on roles in online communities, White et al.[12] 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". In a similar work, Ho et al. [4] presented TopicBlock that combines text and network data for building a taxonomy for a corpus. The LDA model and MMSB models were combined by Nallapati et al. [7] 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. [10] proposed the Bi-LDA model that generates topics from the target languages for paired documents in these very languages. The end-goal of their approach is to classify any document into 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. [2], Yadati and Narayanam [13]). Zhu et al. [14] combine MMSB and text for link prediction and scal it to 44K links.

Ho et al. [5] provide a unique triangulated sampling schemes for scaling mixed membership stochastic block models [1] to hundreds of thousands users based communities. Prem et al. [3] use stochastic variational inference coupled with sub-sampling techniques to scale MMSB like models to hundreds of thousands of users.

None of the works above address the strucutre of the information flow in an online community. Our work is unique in this context as it tries to bridge the gap between community discovery and strucured interaction. We propose a novel modelling scheme that takes into account the network information, user contents as well as structure of the interaction and is scalable to big online forums.

3 Approach

Online forums generally have a specific structure that provides a lot of context to all the interactions occuring among the users. Ignoring this in the analysis makes researhers lose a lot of precious information as we will see in later sections. Here we describe a typeical forum & related structure and the answers that we are looking for.

3.1 Structure in online forums

In an online forum when two users interact in a thread or through a post they probably bring from their own individual point of views or come form possibly different communities. It is valuable to know which topic/community they each belong to in that interaction. When a user U is representing community C out of all communties that he is part of, he tailors his post content accordingly to suit the explicit or implicit community norms. Knowing the style of community specific text content provides a lot of information about the community in general. It also provides information about what role user U plays when he is in community C. In online forums multi-user interactions happen a lot i.e. in a thread a user can post by addressing to another specific user but he is also addressing other users in the thread explicitly or implicitly. Modeling this phenomenon would bring our model closer to reality. This knowledge can be modelled by aggregating users posts acorss a thread, though not across the whole of the forum. We will elaborate on this more in the generative story section. Another interesting property of such structured conversations is that there is an inherent bias towards the thread starter or in turn topic of the thread. It would be interesting to see what insights this knowledge provides given that a model can make use of such an information (This would be very relevant for post-and-response forums in our dataset such as Reddit and Stack Overflow. Right now our graphical model doesn't support this but it would be interesting to see how this can be brought in. It might not be too difficult to do this).

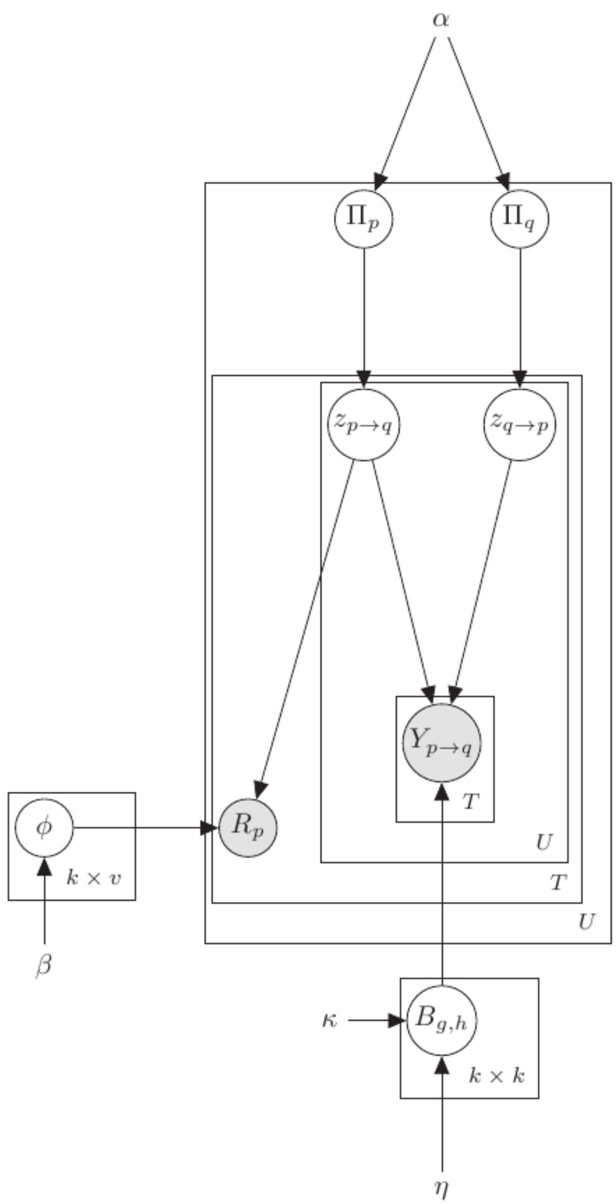

Figure 1: This graphical model takes into account multi-way interaction among users in a thread simultaneously

3.2 Graphical model & generative story

Based on the discussions above our graphical model is designed as 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. This helps us incorporate the knowledge that a user's post is influenced by all the posts of other users present on the thread.

The generative process for the model 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

$$q(z, \Phi, \Pi, B|\Delta_{z_{\to}}, \Delta_{\Phi}, \Delta_{B}, \Delta_{z_{\leftarrow}}, \Delta_{B_{\kappa}}) = \prod_{t}\prod_{p,q}\left(q_1(z^t_{p\to q}|\Delta_{z_{p\to q}}) + q_1(z^t_{p\leftarrow q}|\Delta_{z_{p\leftarrow q}})\right)$$

$$\cdot\prod_{p} q_4(\Pi_p|\Delta_{\Pi_p})\prod_{k} q_3(\Phi_k|\Delta_{\Phi_k})\prod_{g,h} q(B_{g,h}|\Delta_{B_\eta}, \Delta_{B_\kappa}) \tag{3}$$

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

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

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

Equation 5 is the variational lower bound of the log likelihood function which is to be maximized. There are terms like $E_q\left[\sum_{g,h} \log\left(B_{g,h}^{\kappa-1}/\eta^\kappa \Gamma(\kappa) \cdot \exp(-B_{g,h}/\eta)\right)\right]$ which can be obtained by taking derivation of the partition function of the exponential family form of gamma distribution. An effective way to evaluate $E_q\left[\sum_t \sum_p \sum_{w \in R_p^t} \log\left(\prod_{u \in V} (\bar{z}^T \phi_u)^{w=u}\right)\right]$ is by introducing an additional latent variable $\bar{z}_p$ which is a realization of the average $\frac{\sum_{q \in Q} z_{p\to q}}{|Q|}$. So figure 1 will be modifed slightly in future where R_p is drawn from $\bar{z}$ and $\bar{Z}$ is drawn from $z_{p\to q}$. This was suggested by Chong as well as Eric but I finally figured out the equation of the vriational approximation for this. I will update the equations once I have coded it and verified..

4 Datasets

We analyse three different forums: 1) Cancer forum, 2) Stack Overflow, and 3) Reddit. The three datasets mentioned above represent three different sets of online gatherings which helps us genearlize pur claims.

4.1 Cancer Forum

The cancer forum is an online forum where users discuss about their cancer treatment and any thing else under the sun. Here again the conversation happens in a structured way where users post their responses on a thread by thread basis. Users also call each other by their names (or nick-names) while posting in many cases. This forum has around 3000 users and 10,000 threads, and a user on average posts around 120 words in a post.

4.2 Stack Overflow

This is an online forum where users ask and answer technical troubles. It is a typical online forum where users reply to each other in a threaded structure. Based on the response the replies are voted up and down by other users. This voting score is used in our prediction tasks later. Shriphani will provide the exact statistics of this dataset as soon as the crawl is done. We will have most likely have around 1/2 a million users in this set.

4.3 Reddit

Reddit is an online trend spotting website where users post interesting articles, news, stories, links etc. and a discussion ensues. Users can upvote or downvote any reply or posts. The converstaions happen in a threaded structure as described in our generative story. The upvotes are used later by our model for prediction task. This is again an ongoing crawl but we should have atleast 200K+ users here. Shriphani please provide the latest numbers soon

5 Experimental Setup and Evaluation

We perform user-user link prediction tasks as well as user survival predicition in forums besides reporting perplexity and log-likelihood convergence on held-out test set. We also analyse and the user-communtiy vectors Π as defined in section 3 to show-case interesting insight that the newly discovered user-communities provide.

5.1 Link prediction

We hold out some users randomly in the dataset and predict their likelihood of interaction with other users based solely on the content of their posts in the forum and the parameters of the model learnt in the training phase. We take the top K (1, 5 and 10) users that the model predicts for every held-out user and report the overall precision and recall in the top-K set. This will have to be done for a specific snapshot of the forum because not many users will have a overlapping posting history accross their whole stay on the forum.

5.2 Cancer forum user survival prediction

Users in cancer forum posts different messages at different phases of theor cancer. It has been seen that in while users are certain cancer phase they tend to post more often and regularly. The intuition behind this prediction task is to exploit this pattern. We use the variational parameter of the $\bar{Z}_{p\rightarrow}$ to get the topic composition of a users posts in the forum at one particular snapshot of time e.g. every month or every two months. We combine the features shown to be useful for such an analysis by Wen et al. [11] with the variational parameters to predict whether the user will post in the coming period of time e.g. within the next month or two months.

6 Results

7 Conclusion & Future Work

1) Currently our model just picks up one signal for the network component (MMSB part) i.e. in other words we are just modelling one type of interaction or just one graph, but as we did for the SEI model (tensor model) there are multiple types of networks/graphs/interactions. Future work can incorporate this.
2) Adding temporal dimension to this model would be a very interesting idea. E.g. how threads evolve over time, or how user behavior changes, or how new communities emerge in the forum etc.

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] 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.

[3] Prem Gopalan, David M. Mimno, Sean Gerrish, Michael J. Freedman, and David M. Blei. Scalable inference of overlapping communities. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Lon Bottou, and Kilian Q. Weinberger, editors, *NIPS*, pages 2258–2266, 2012.

[4] 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.

[5] Qirong Ho, Junming Yin, and Eric P. Xing. On triangular versus edge representations — towards scalable modeling of networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Lon Bottou, and Kilian Q. Weinberger, editors, *NIPS*, pages 2141–2149, 2012.

[6] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347, May 2013.

[7] 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.

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

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

[10] 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.

[11] Miaomiao Wen, Zeyu Zheng, Hyeju Jang, Guang Xiang, and Carolyn Rose. Extracting events with informal temporal references in personal histories in online communities. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: Short Papers*, ACL '13, Stroudsburg, PA, USA, 2013. Association for Computational Linguistics.

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

[13] 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.

[14] Y. Zhu, X. Yan, L. Getoor, and C. Moore. Scalable Text and Link Analysis with Mixed-Topic Link Models. *ArXiv e-prints*, March 2013.