Unified Modeling of User Activities on Social Networking Sites

Himabindu Lakkaraju
IBM Research - India
Manyata Embassy Business Park
Bangalore, Karnataka - 560045
klakkara@in.ibm.com

Angshu Rai
IBM Research - India
Manyata Embassy Business Park
Bangalore, Karnataka - 560045
angshu.rai@in.ibm.com

Abstract

Social networking sites like Facebook and Twitter are teeming with users and the content posted by them. Several activities like friendship/followership, authoring, commenting on, liking, resharing/retweeting posts typically occur on these sites. In this paper, we make an attempt at the unified modeling of various such activities on social networking sites. We propose a novel joint latent factor model, Latent User Preference Model (LUPM), which combines the predictive power of multiple dyadic relations and text content using block and topic models coupled through a common latent representation for the users and posts. We further experiment with real world Twitter and Facebook datasets in order to understand the empirical significance of such unified modeling.

1 Introduction

Since the advent of social networking sites like Facebook and Twitter, an unprecedented number of people have started using them as a mode of communication and a medium to share thoughts. Millions of users registered with these sites interact with each other. The diversity of the users and content available on these sites is what makes the activity on these sites interesting. Users on these sites may engage in several activities like friendship/followership, authoring posts, commenting on posts etc., however there would still exist certain underlying patterns. To illustrate, let us consider a user named Sally who is interested in movies and sports. Her interests are most likely to manifest into actions in the form of postings and also followership. Sally may post something about movies she has watched or wants to watch, follow a movie star. In all these actions, there is an underlying pattern that is motivating Sally to perform these activities which is her interest in movies. This underlying interest has influenced multiple actions of Sally. Only a holistic analysis of all her actions can facilitate understanding of such underlying interests.

Most of the work on social networking sites till date analyzes either the content [3], [4] or the network links. In this work, we attempt the joint modeling of various user activities on social networking sites by proposing a novel joint latent factor model, Latent User Preference Model (LUPM), which combines the predictive power of multiple dyadic relations and text content using mixed membership stochastic block models and topic models coupled through a common latent representation for the users and posts. Experimentation on real world Facebook and Twitter data demonstrates the efficacy of the proposed approach on link prediction and content personalization tasks.

2 Our Approach

In this section, we present joint latent factor model, Latent User Preference Model (LUPM). This model employs stochastic block models for capturing the interactions between users and posts and the relationships among users, and topic models for analzying the textual content. LUPM integrates all these signals from various user activities on social networks, thus facilitating a holistic view of the user behavior on social media. We use the following notations through out our description - $\mathcal{U}$

denotes the set of users, $\mathcal{P}$, the set of posts. For each user $u \in \mathcal{U}$, $\mathbf{c}_u^U$ and $\mathbf{x}_u$ denote the text content and structured attributes (e.g., gender) in the user profile. For each post $p \in \mathcal{P}$, $\mathbf{c}_p^P$ and $\mathbf{y}_p$ denote the text content and structured attributes, (e.g., hasLink). Further, for each dyad of users $(u, v) \in \mathcal{U} \times \mathcal{U}$, $f_{u,v}$ denotes the friendship / follower relationships between the user dyad (u, v). Also, $i_{u,p}$ denotes encoding of interactions (replying, sharing etc.) between the dyads $(u, p) \in \mathcal{U} \times \mathcal{P}$. Similarly, $a_{u,p}$ denotes the post authorship relation between the user u and post p. The generative process for

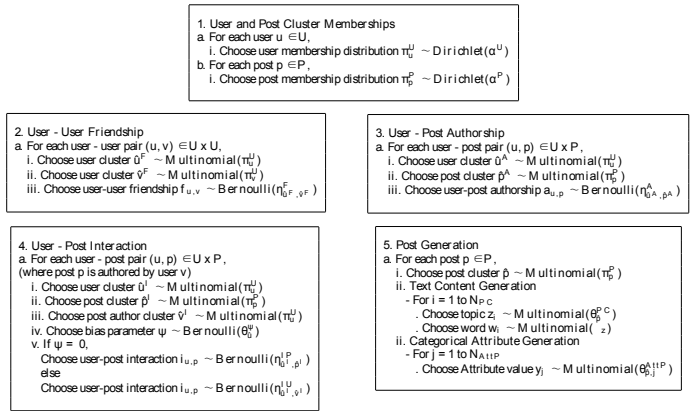

Figure 1: Latent User Preference Model - Generative Process

LUPM is highlighted in Figure 1. In order to facilitate easy understanding, we highlight the different parts of the generative process with the appropriate titles. At the heart of this model lie Mixed Membership Stochastic Block models for modeling the dyadic interactions [1] and the topic model Latent Dirichlet Allocation [2] for modeling the textual content. Now, we explain the generative process in detail :

User and Post Cluster Memberships : The flexibility provided by this model can be attributed to the mixed membership provision. According to this, each user has a certain probability with which his/her behavior (activities and relationships) can be attributed to a particular cluster. Each user u is associated with a vector π_u^U which encodes these probabilities. $\pi_{u,g}^U$ denotes the probability that the behavior of the user u is governed by cluster g in the case of a particular user activity. Similarly, each post is associated with a mixed membership vector π_p^P. These membership vectors govern the behavior (in the case of users) and the content (in the case of posts) across all the dyadic relationships. α^U and α^P correspond to the symmetric dirichlet priors of all the user and post membership vectors respectively.

User - User Friendship : This corresponds to an important dyadic user - user relationship in social networking websites. For each user - user dyad of the form (u, v), the appropriate user cluster is sampled from the user specific membership vectors for both u and v. η^F is a $K^{\mathcal{U}}$ x $K^{\mathcal{U}}$ ($K^{\mathcal{U}}$ being the number of user clusters) dimensional matrix comprising of the probabilities of relationships between pairs of user clusters. Precisely, $\eta^F_{\hat{u},\hat{v}}$ denotes the probability of the existence of a relationship between a pair of users assigned to the user clusters $\hat{u}$ and $\hat{v}$ respectively. Further, we assume that each element of the matrix $\eta^F_{\hat{u},\hat{v}}$ is drawn from a symmetric dirichlet prior, $Dirichlet(\alpha^F)$. The matrix f encodes the observed relationships between various user pairs i.e if a pair of users u and v are friends, then the corresponding matrix entry $f_{u,v}$ is 1, otherwise it is a 0.

User - Post Authorship : This represents the dyadic authorship relation between the users and posts. The matrix a captures this relationship between the users and the posts - if a user u has authored a post p, then the element $a_{u,v}$ is 1, it is a 0 otherwise. Also, the matrix η^A encodes the probabilities of the existence of authorship relation between the user clusters and the post clusters. The generative process is similar to the user - user friendship.

User - Post Interaction : This relationship forms a very crucial part of the model. As motivated in the introduction, a user would typically interact(respond/share etc.) with a post if either the content of the post is of interest or the person who has authored the post is of interest. In order to model this, we introduce a bias parameter ψ. For each user - post pair, the user and post clusters, $\hat{u}^I$ and $\hat{p}^I$, corresponding to the interaction are chosen. In addition to these, the user cluster corresponding to the author of the post $\hat{v}^I$ is also chosen. Further, a bias parameter ψ is drawn from a binomial

distribution conditioned on the user cluster $\hat{u}^I$, $\theta^{\psi}_{\hat{u}^I}$. If this parameter is 0, then the interaction between the user and post is governed by $\hat{u}^I$ and $\hat{p}^I$, otherwise, the interaction depends on the user clusters $\hat{u}^I$ and $\hat{v}^I$. This ensures that the probability with which a user is biased towards content of the posts versus the authors of the posts is determined by the cluster to which the user belongs to. The parameters η^{IP} and η^{IF} are analogous to the η parameters described above. Further, the matrix i captures the interactions between the users and posts, if a user u has interacted (replied to, shared, favorited etc.) with the post p, then the element $i_{u,p}$ is assigned a value 1, else it is a 0.

Post Generation : This part of the model is motivated by one of the most widely known topic models Latent Dirichlet Allocation ([2]). However, instead of relying on document level topic mixtures, our model assumes that each post is assigned a post cluster. This post cluster determines the topic distributions of the post and also the attributes of the post. Each post cluster is associated with a multinomial distribution over topics $\theta^{PC}_{\hat{p}}$ and the text content of each post follows this distribution. The topic distribution ϕ is drawn from Dirichlet(β). Further, the categorial attributes (eg. hasLink, isPhoto etc.) also depend upon the cluster level multinomial distribution $\theta^{AttP}_{\hat{p}}$.

3 Inference

Here we perform approximate inference using collapsed gibbs sampling [**?**]. Due to space constraints, we refrain from presenting the complete derivation, instead we present only the update equations. Note that in all the update equations that we present here, the instance which is sampled is omitted from the counts which determine the sampling distribution. The following are the update equations :

User - User Friendship:

$$P(\hat{u}^F = l|\hat{v}^F, f) \propto (n^{\mathcal{U}}_{u,l} \times \alpha^{\mathcal{U}}) + (n^{F}_{l,\hat{v}^F,f_{u,v}} + \alpha^F)$$
$$P(\hat{v}^F = m|\hat{u}^F, f) \propto (n^{\mathcal{U}}_{v,m} + \alpha^{\mathcal{U}}) \times (n^{F}_{\hat{u}^F,m,f_{u,v}} + \alpha^F)$$

$n^{\mathcal{U}}_{u,l}$ denotes the number of times a user u is assigned the user cluster l across all the interactions and relationships. $n^{F}_{l,\hat{v}^F,f_{u,v}}$ denotes the number of times that the friendship relations directed from user cluster l to $\hat{v}^F$ have the value $f_{u,v}$. The notations used for sampling the user cluster corresponding to the user v can be explained analogously.

User - Post Authorship:

$$P(\hat{p}^A = r|\hat{u}^A, a) \propto (n^{\mathcal{P}}_{p,r} + \alpha^{\mathcal{P}}) \times (n^{A}_{\hat{u}^A,r,a_{u,p}} + \alpha^A)$$

where $n^{\mathcal{P}}_{p,r}$ denotes the number of times a post p is assigned the post cluster r across all the interactions. Further, $n^{A}_{\hat{u}^A,r,a_{u,p}}$ denotes the number of times authorship relations directed from user cluster $\hat{u}^A$ to the post cluster r take the value $a_{u,p}$.

User - Post Interaction:

$$P(\hat{u}^I = l|\hat{v}^I, \hat{p}^I, \psi, i) \propto (n^{\mathcal{U}}_{u,l} + \alpha^{\mathcal{U}}) \times (n^{IP}_{l,\hat{p}^I,i_{u,p}} + \alpha^{IP}) \text{ if } \psi = 0,$$
$$(n^{\mathcal{U}}_{u,l} + \alpha^{\mathcal{U}}) \times (n^{IU}_{l,\hat{v}^I,i_{u,v}} + \alpha^{IU}) \text{ otherwise}$$

The clusters corresponding to the post author v and the post u can be sampled in a similar manner. The bias parameter ψ can be sampled from the following distribution :

$$P(\psi = 0|\hat{u}^I, \hat{v}^I, \hat{p}^I) \propto (n^{\psi}_{\hat{u}^I,0} + \alpha^{\psi}) \times (n^{IP}_{l,\hat{p}^I,i_{u,p}} + \alpha^{IP})$$
$$P(\psi = 1|\hat{u}^I, \hat{v}^I, \hat{p}^I) \propto (n^{\psi}_{\hat{u}^I,1} + \alpha^{\psi}) \times (n^{IU}_{l,\hat{v}^I,i_{u,v}} + \alpha^{IU})$$

where $n^{\psi}_{\hat{u}^I,0}$ corresponds to the number of times ψ has been assigned the value 0 w.r.t user cluster $\hat{u}^I$. Rest of the notations are analogous to those described so far.

Post Generation:

$$P(\hat{p} = s|\mathbf{z}_p, \mathbf{y}_p) \propto (n^{\mathcal{P}}_{p,s} + \alpha^{\mathcal{P}}) \times \prod_{i=1}^{N_{PC}} (n^{PC}_{\hat{p},z_i} + \alpha^{PC}) \times \prod_{j=1}^{N_{AttP}} (n^{AttP}_{\hat{p},y_j} + \alpha^{AttP})$$

The topic for each word in the content of a post p given the post cluster p and other relevant parameters is sampled using

$$P(z_i = t|\hat{p}, \cdot) \propto (n^{PC}_{\hat{p},t} + \alpha^{PC}) \times \frac{\left(n^t_v + \beta\right)}{\left(\sum_{r=1}^{V}(n^t_r + \beta)\right)}$$

$n^{PC}_{\hat{p},t}$ corresponds to the number of words in the documents sampled from post cluster $\hat{p}$ that are assigned to the topic t. Let the word in the i^{th} position of the post p correspond to the v^{th} word of the vocabulary, then, n^t_v corresponds to the number of times word v is assigned the topic t.

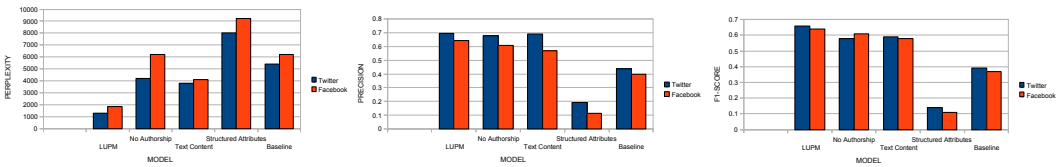

Figure 2: a. Perplexity Evaluation b. User Friendship/Followership Link Prediction c. User-Post Interaction Prediction

4 Experimental Results

In this section we discuss in detail the experiments that we carried out using the proposed model on real world social media datasets extracted from Facebook (data comprising of 257 users and 2441 posts was collected for a span of 3 weeks) and Twitter (data comprising of 3230 users and 346K posts spanning a period of about 2 months). We present here three different kinds of experimental results - Perplexity evaluation, User friendship/followership link prediction and User post interaction prediction. For all of our experiments, we ran the collapsed gibbs sampling algorithm for 1000 iterations. The hyper-parameters were initialized as: $\alpha^{\mathcal{U}} = 50/K^{\mathcal{U}}$, $\alpha^{\mathcal{P}} = 50/K^{\mathcal{P}}$, $\alpha^{AttP} = 50/K^{AttP}$, $\alpha^{PC} = 50/K^{T}$, $\alpha^{A} = \alpha^{F} = \alpha^{IP} = \alpha^{IU} = \alpha^{\psi} = \beta = 0.1$.

Ablations We tried to analyze how the predictions would be affected if we eliminate the user - authorship block model from LUPM, we denote this using No authorship label in our results in Figure 2. Further, we consider only the text content in the post generation process (ignoring the generation of post attributes), we denote this variation using the label Text Content in our results. Lastly, we consider only the structured attributes of posts without the text content and denote this using the label Structured Attributes.

Perplexity Evaluation : Perplexity is one of the most widely employed empirical measure to detect how well a given model will be able to generalize to the test data. The perplexity, used by convention in language modeling, is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood. In order to compute model perplexity, we consider the dataset spanning the last one week of data. Figure 2a. depicts these measurements on Twitter and Facebook data for our model LUPM, its ablations and baseline, a variant of Labeled-LDA [3]. It can be seen that our approach gives the least perplexity in both the cases.

Predicting User Friendship/Followership Links : We carried out two different kinds of evaluations relating to this. Firstly, we masked about 10% of the friendship/followership links in the available data and checked if the algorithm could predict the same. This is essentially a precision based task. The accuracy numbers are highlighted in Figure 2b. It can be seen that the joint model outperforms the baseline and other ablations. Secondly, we collected the user-user pairs which had no friendship/followership relation between them and obtained predictions from the algorithm. If the algorithm predicted that there should be a link between a particular pair, we treated it as a recommendation and let the user evaluate this recommendation. We collected inputs from about 20 users for a total of 92 recommendations and 73 out of the 92 recommendations provided by the joint model were evaluated as relevant.

Predicting User-Post Interactions : For this part of the experimentation, we used a test set on which the prediction of the interactions estimated by the model were evaluated. In case of Twitter data, we set aside 80K posts. These were the posts generated during the last 2 weeks of the 3 month span over which the data was collected. For the Facebook dataset, we used the data which emerged in the various user news feeds for the last 6 days over a period of 3 weeks which accounted for about 678 posts. Apart from experimenting with the joint model and the baseline, we also tried to evaluate the impact of the individual signals on the predictions. For each user, we collected all the posts that he/she would come across during the appropriate period of time and also the posts that the user actually interacted with (commented/shared/favorited etc.). We computed the F1-measure in order to determine how well the proposed approach was able to predict such interactions. The results are highlighted in 2c. As can be seen, the joint model, LUPM, does better than all the other models. It is interesting to note that, in case of Facebook data, excluding the authorship signal from the model improves the F1-score slightly, this was due to the fact that some users posted content on concepts pertaining to movies, but when their friends authored some of the content relevant to the concept, they chose not to interact with it. Owing to this kind of behavior, the precision slightly increased on excluding this signal from the model.

Acknowledgements

The authors would like to thank Srujana Merugu for insightful discussions and helpful comments.

References

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. 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. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing microblogs with topic models. In *International AAAI Conference on Weblogs and Social Media*, 2010.

[4] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.