Smart News Feeds for Social Networks Using Scalable Joint Latent Factor Models

Himabindu Lakkaraju
IBM Research - India
klakkara@in.ibm.com

Angshu Rai
IBM Research - India
angshu.rai@in.ibm.com

Srujana Merugu
IBM Research - India
srujanamerugu@in.ibm.com

ABSTRACT

Social networks such as Facebook and Twitter offer a huge opportunity to tap the collective wisdom (both published and yet to be published) of all the participating users in order to address the information needs of individual users in a highly contextualized fashion using rich user-specific information. Realizing this opportunity, however, requires addressing two key limitations of current social networks: (a) difficulty in discovering relevant content beyond the immediate neighborhood, (b) lack of support for information filtering based on semantics, content source and linkage, which can be interpreted in terms of predicting user-post relevance with better recall and precision respectively.

We propose a scalable framework for constructing *smart news feeds* based on predicting user-post relevance using multiple signals such as text content and attributes of users and posts, and various user-user, post-post and user-post relations (e.g., friend, comment, author relations). Our solution comprises of two steps where the first step ensures scalability by selecting a small set of user-post dyads with potentially interesting interactions using inverted feature indexes. The second step models the interactions associated with the selected dyads via a joint latent factor model, which assumes that the user/post content and relationships can be effectively captured by a common latent representation of the users and posts. Experiments on a Facebook dataset using the proposed model lead to improved precision/recall on relevant posts indicating potential for constructing superior quality news feeds.

1. INTRODUCTION

The ever-increasing participation of authoritative news sources on social networks coupled with rich multi-media support, and flexible communication protocols have resulted in social networks such as Facebook/Twitter being well-positioned to become the dominant acquisition and dissemination systems for both generic and personal information. Most users, however, still use a combination of sources as news sites, search engines, and Q & A forums even though the relevant information resides somewhere on Facebook/Twitter and can be delivered more effectively by taking into account user demographics, network linkage and fine-grained historical activity. This is primarily due to two reasons that emerge in user studies [5]: First, it is currently non-trivial to discover all the relevant information or sources in a social network beyond the immediate social graph. Second, the current news feeds in social networks are based on the immediate social graph with little customizability. Expansion of one's social graph to include all potentially relevant sources would, thus, result in the user's feed being inundated with a lot of irrelevant content that has to be manually perused.

Addressing the above limitations in social networks in an effective way can help build highly useful applications such as (i) smart news feeds comprising of a mix of highly relevant generic and personal information from all over the network, (ii) automatic generation of relevant responses to a query from existing network content, and (iii) question-answering on network based on intelligent routing of queries to expert users. The key technical challenge is to design scalable techniques that can combine a large variety of sparse, high dimensional signals, such as text content and attributes of posts and users, and dyadic user-user, post-post, user-post relations (e.g., network linkage, authorship, commenting activity) to predict other relationships of interest, e.g., the relevance of a post to a user or to another post.

Currently, there exists related work in the area of personalized news recommendation [4] and social network-based search [6] where the relevance of a post to a user is modeled in terms of the structured user-post attributes and user-user (activity or linkage) correlation . Of these, the techniques based on discriminative models, require substantial feature engineering effort in addition to handling missing observations, while those based on generative models are not very scalable and handle a single dyadic relation.

In the current work, we consider the problem of constructing a smart news feed by modeling user-post relevance. The novelty of our approach lies in (i) ensuring scalability of the generative model for user-post interactions by conditioning it on a selection variable, which can be computed fast using inverted feature indexes in a prior step, and (ii) combining the predictive power of multiple dyadic relations and text content using block and topic models coupled using a common latent representation for the users and posts. Section 2 provides a formal problem statement while Section 3 describes the solution approach.

2. PROBLEM STATEMENT

Let $\mathcal{U}$ denote the set of users, $\mathcal{P}$, the set of posts. For each user $u \in \mathcal{U}$, let $\mathbf{c}_u^U$ and $\mathbf{x}_u$ denote the text content and demographic attributes (e.g., gender) in the user profile. Similarly, for each post $p \in \mathcal{P}$, let $\mathbf{c}_p^P$ and $\mathbf{y}_p$ denote the text content and structured attributes, (e.g., hasLink). Further, for each dyad of users $(u, v) \in \mathcal{U} \times \mathcal{U}$, let $\mathbf{r}_{u,v}^{UU}$ denote a vector encoding various relationships between the user dyad (u, v), e.g., friend, follower, etc. Similarly, let $\mathbf{r}_{p,q}^{PP}$ and $\mathbf{r}_{u,p}^{UP}$ denote encoding of relationships between the dyads $(p, q) \in \mathcal{P} \times \mathcal{P}$ and $(u, p) \in \mathcal{U} \times \mathcal{P}$. Given observations on user and post-specific properties, and (possibly incomplete) user-user, post-post, user-post relationships, the goal is to predict user-post relevance (or in general, some fine-grained user-post interaction), in order to obtain all the relevant posts for each user.

3. SOLUTION APPROACH

We address the above problem using a two step approach. The first step is motivated by scalability concerns and involves selecting a small set of dyads with potentially interesting interactions using inverted feature-based indexes. The second step assumes the first step selection variables to be fully observed, and models the dyadic interactions, post and user content using a joint latent factor model. To effectively capture the key post content aspects, we use the labeled-LDA model similar to the one employed in [5]. Each of these components

is discussed below.

Selection of Interesting Dyads. We choose a set of easy-to-compute, predictive features for the relevant dyad types (user-user, post-post, user-post). For instance, in the case of user-post dyads, these include network distance between user and post author, immediate neighborhood size of post author, etc. The observed data is preprocessed to create maps from (user, feature, feature-values) to matching posts. For each feature, a threshold is identified such that the feature-threshold predicate (e.g., feature value $<$ threshold) provides high recall and good precision for user-post pairs with interesting interactions. These predicates are then combined disjunctively to yield the desired high filter. This process is akin to filtering using inverted word/phrase to document maps in search engines prior to estimating the query-document relevance with the main difference being the use of general features instead of words/phrases.

Joint Latent Factor Model. The joint model in the second step is based on a key assumption that the dyadic (user-user, user-post, post-post) interactions and user/post text content can be explained using a compact latent representation of users and posts along with the observed user/post attributes. This latent representation not only provides a principled way of integrating information from multiple dyadic relations and text content, but also helps in avoiding issues arising due to high dimensionality and sparsity. The latent representation for a user u and a post p takes the form of a mixed membership across multiple user and post clusters and denoted by π_u^U and π_p^P respectively. The generative process can be briefly summarized as follows:

$$\pi_u^U \sim Dir(\alpha^U),\ \pi_p^P \sim Dir(\alpha^P),\ \forall u \in \mathcal{U}, \forall p \in \mathcal{P},$$
$$z_{u,v}^{UUS}, z_{u,v}^{UUT}, z_{u,p}^{UP}, z_u^{UC} \sim Mult(\pi_u^U),\ \forall u, v \in \mathcal{U}, \forall p \in \mathcal{P},$$
$$z_{p,q}^{PPS}, z_{p,q}^{PPT}, z_{p,u}^{PU}, z_p^{PC} \sim Mult(\pi_p^P),\ \forall p, q \in \mathcal{P}, \forall u \in \mathcal{U},$$
$$\mathbf{r}_{u,v}^{UU} \sim BM^{UU}(\eta^{UU}, z_{u,v}^{UUS}, z_{v,u}^{UUT}, \beta^{UU}, \mathbf{x}_u, \mathbf{x}_v),\ \forall u, v \in \mathcal{U},$$
$$\mathbf{r}_{p,q}^{PP} \sim BM^{PP}(\eta^{PP}, z_{p,q}^{PPS}, z_{q,p}^{PPT}, \beta^{PP}, \mathbf{y}_p, \mathbf{y}_q),\ \forall p, q \in \mathcal{P},$$
$$\mathbf{r}_{u,p}^{UP} \sim BM^{UP}(\eta^{UP}, z_{u,p}^{UP}, z_{p,u}^{PU}, \beta^{UP}, \mathbf{x}_u, \mathbf{y}_p),\ \forall u \in \mathcal{U},\ \forall p \in \mathcal{P},$$
$$\mathbf{c}_u^U \sim TM^C(\theta^C, z_u^{UC}, \mathbf{x}_u),\ \forall u \in \mathcal{U},$$
$$\mathbf{c}_p^P \sim TM^P(\theta^P, z_p^{PC}, \mathbf{y}_p),\ \forall p \in \mathcal{P}.$$

Here, BM and TM refer to block models and topic models respectively. The variables α^U, α^P denote the hyper-parameters corresponding to the cluster membership priors, while $z_{u,\cdot}^{\cdot}$ and $z_{p,\cdot}^{\cdot}$ denote realizations of user and post cluster labels for different dyads, and z_u^{UC} and z_p^{PC} denote the ones that influence the user/post text content. The dyad specific cluster labels along with the block model parameters η, the observed user and post attributes $(\mathbf{x}_u, \mathbf{y}_p)$, and attribute coefficients β generate the dyadic interactions via generalized linear models (GLM) as in [2](e.g., logistic regression for binary interactions). Similarly, the cluster labels z_u^{UC}, z_p^{PC} along with the topic model parameters θ and user/post attributes $(\mathbf{x}_u, \mathbf{y}_p)$ generate the user profile and post content using an LDA model [3]. Estimation of parameters and latent variables in the joint model is done using an approximate Gibbs sampling algorithm assuming default interactions for the dyads not selected in the first step.

Aspect LDA Model for Posts. We observe that posts can be associated with few key aspects that are highly predictive of user-post interactions. These include: (a) scope - personal/non-personal, (b) message type - sentiment/question/statement, (c) concept- science/sports, etc. Of these, "personal", "sentiment", "question" and specific instances of "concepts" can be readily defined in terms of word distributions while the other aspect choices tend to be defined by their complementary relation and hence, we learn a LDA model [3] with these topics seeded with a few exemplary words.

4. EXPERIMENTAL RESULTS

Using a Facebook application that users can sign up for, we constructed a dataset comprising of user friendship linkage, user-post authorship, commenting activity, user profiles and detailed post content (text/attributes such as message type-status/video/link). We considered a subset of this data with 257 users, 2441 posts and 1158 known friendship links. To evaluate the proposed model, we assume that one or more comments by a user on a viewed post is a proxy for the user-post relevance that we seek to predict. Table 1 shows prediction quality for 4075 pairs of user and viewed posts (i.e., user is friend of post author) from the collected data using different approaches: baseline Facebook feed, proposed joint modeling approach, as well as separately using the different predictive signals. The models were trained with 6 user clusters, 7 post clusters, 7 topics, and inference algorithm run for 500 iterations. The results point to the efficacy of the joint latent factor model in filtering out irrelevant posts while retaining most of the relevant ones. We also performed an evaluation of *smart feeds* constructed from unseen posts (i.e., not authored by friends) using the joint model on a small set of users and obtain a precision in the range 61.11% indicating that the joint model enables discovery of new content that is not in the immediate neighborhood. We further analyzed the joint model parameters and latent clusters to identify interesting correlations. For instance, user cluster 4 exhibits a high tendency to comment on posts in post clusters 4 as shown in Figure 1, which is primarily about sports topics and authored by users in user clusters 1 and 4. More details on evaluation are provided at [1].

	Baseline	Joint Model	Text Content	Structured Attributes	Past Comments
Precision	0.072	0.3659	0.3078	0.1915	0.1917
Relative Recall	1.0	0.8851	0.7331	0.4726	0.1418
F-measure	0.1343	0.5178	0.4334	0.2725	0.1630

Table 1: User-post relevance prediction (3 fold CV).

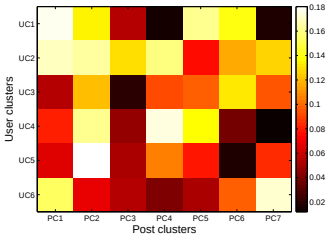

Figure 1: Average probability of commenting.

5. CONCLUSION AND FUTURE WORK

Preliminary results using the joint latent factor model based approach for estimating user-post relevance are quite promising, but further experimentation with different cluster parameters, inference schemes and larger data sets needs to be performed. We also plan to apply this methodology to other applications such as question routing and auto-response generation in the near future. The dynamic nature of social network data also makes it critical to study incremental inference algorithms and temporal trends. Other potential extensions include using hierarchical block models for dyadic relations and syntactic topic models for text content.

6. REFERENCES

[1] https://researcher.ibm.com/researcher/view.php?person=in-klakkara.
[2] D. Agarwal and S. Merugu. Predictive discrete latent factor models for large scale dyadic data. In *KDD*, pages 26–35, 2007.
[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.
[4] A. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In *WWW*, pages 271–280, 2007.
[5] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.
[6] Z. Wen and C. Lin. On the quality of inferring interests from social neighbors. In *KDD*, pages 373–382, 2010.