TEM: A Novel Perspective to Modeling Content on Microblogs

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

Hyung-Il Ahn
IBM Research - Almaden
hiahn@us.ibm.com

ABSTRACT

In recent times, *microblogging* sites like Facebook and Twitter have gained a lot of popularity. Millions of users world wide have been using these sites to post content that interests them and also to voice their opinions on several current events. In this paper, we present a novel non-parametric probabilistic model - Temporally driven Theme Event Model (TEM) for analyzing the content on microblogs. We also describe an online inference procedure for this model that enables its usage on large scale data. Experimentation carried out on real world data extracted from Facebook and Twitter demonstrates the efficacy of the proposed approach.

Categories and Subject Descriptors

H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

social media, nonparametric topic models, scalable inference

1. INTRODUCTION

Microblogging sites such as Facebook and Twitter are housing huge volumes of user generated content. Analyzing the content on these sites poses new challenges and opportunities because of the scale of the data and also the nature of the content. The content on these microblogging sites can be seen as a consequence of an intricate interplay between themes of individual interests, events occuring across the world and their temporal dynamics. To illustrate, let us consider a microblogger who is an avid football fan, it is most likely that this user will author several posts related to football. Thus, one of his themes of interest is 'football'. Further, when a particular event relevant to this theme occurs for ex. FIFA worldcup, it is likely that this user is going to post about this event. Thus, the postings on microblogs can be seen as a combination of theme related content and user responses to events occuring at a particular instant in time. Characterizing content on microblogs can benefit from capturing the interplay of these two aspects.

Most of the work involving analysis of the content on microblogs[4] relies on traditional text analysis techniques such as Latent Dirichlet Allocation[2]. A few variants of traditional algorithms have been proposed to analyze microblogs[6]. However, these methodologies require meta-information such as hashtags which are specific to certain microblogging sites. In this work, we attempt to overcome these limitations by proposing a rich non-parametric probabilistic model TEM that effectively captures the generation of the content on microblogs. Further, we describe an online gibbs sampling based inference algorithm for the proposed model and experiment with real world datasets in order to study the effectiveness of the proposed methodology.

2. OUR APPROACH

In this section, we discuss the approach that we propose for modeling the generation of the content in microblogs. We assume that each post is generated by a single underlying theme (for ex. sports) and a single underlying event (for ex. FIFA world cup). We employ Dirichlet process priors[7] for modeling the generation of the themes and events. In order to provide more weightage to the most recent user preferences and account for clustering of posts based on their temporal proximity, we incorporate time decaying kernels in the dirichlet process priors. The model that we describe can be essentially broken down into three components - Identifying underlying theme of a post; Identifying the event associated with the post; Modeling each post as a theme/event mixture. We describe each of these components below.

Identifying underlying theme : One of the most significant factors influencing the generation of a post is the *theme.* To illustrate, posts such as 'watching a football match today', 'just finished playing football match' carry an underlying theme which is 'football'. When we see posts corresponding to such underlying theme from a user, it is easy to conclude that the user is interested in football. Thus, themes are governed by individual user preferences. Further, user preferences tend to vary with time. In order to assimilate these aspects into our approach, we model the probability that the post p authored by user u is generated by the theme k - $P_Z(z_{u,p} = k|z_{u,1:(p-1)}, \alpha_Z)$ as -

$$P_Z(z_{u,p} = k|z_{u,1:(p-1)}, \alpha_Z) \propto \begin{cases} \sum_{\delta=1}^{\Delta} e^{-\delta/\lambda} n_{u,k,t-\delta} & \text{if k}\leq\text{K} \\ \alpha_Z & \text{if k=K+1} \end{cases} \quad (1)$$

$n_{u,k,t}$ is the number of posts authored by user u at time instant t which have been assigned theme k. α_Z is a positive real value which ensures that a new theme can be sampled if necessary. K denotes the number of themes and the value of this parameter is not fixed. This essentially gives the model flexibility to dynamically increase the number of themes whenever need arises. Note that while dealing

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

with data sources like Facebook and Twitter, it is extremely tough to assign a value to the parameters like the number of themes and events, hence we resort to a non-parametric prior.

Identifying associated event : In addition to a broad underlying theme, microtext based postings are usually associated with an event. For instance, though posts like 'following FIFA worldcup', 'excited to watch NFL finals today' relate to the same underlying theme, the events that they refer to are different. As discussed above, the theme of a post is governed by user interest, however, the event which a post refers to is governed by other posts from across the globe within a given period of time. Thus, we model the probability that the post p authored by user u is generated by the event q - $P_E(e_{u,p} = q|e_{u\in \mathrm{U},1:(p-1)}, \alpha_E)$ where $\mathcal{U}$ is the set of all the users as -

$$P_E(e_{u,p} = q|e_{u\in \mathrm{U},1:(p-1)}, \alpha_E) \propto \begin{cases} \sum_{\delta'=1}^{\Delta'} e^{-\delta'/\lambda'} n_{q,t-\delta'} & \text{if q}\leq\text{Q} \\ \alpha_E & \text{if q=Q+1} \end{cases} \tag{2}$$

$n_{q,t}$ is the number of posts authored at time instant t that have been assigned the event q. Note that this count is different from the count used for theme identification.

Table 1: Generative Process of Theme Event Model

```
1. For each theme $k$,
      Choose $\phi_k^Z \sim Dir(\beta^Z)$
2. For each event $q$,
      Choose $\phi_q^E \sim Dir(\beta^E)$
3. For each post $p$ authored by user $u$,
      a. Choose theme $z_{u,p} \sim P_Z(z_{u,p}|z_{u,1:(p-1)}, \alpha_Z)$
      b. Choose event $e_{u,p} \sim P_E(e_{u,p}|e_{u\in \mathrm{U},1:(p-1)}, \alpha_E)$
      c. Choose $\pi_{u,p} \sim Beta(\alpha_{mix})$
      d. For each word index $n$ of post $p$
         i. Choose $r_{u,p,n} \sim Mult(\pi_{u,p})$
         ii. If $r_{u,p,n} = 1$, choose $w_{u,p,n} \sim \phi^Z_{z_{u,p}}$
             else if $r_{u,p,n} = 2$, choose $w_{u,p,n} \sim \phi^E_{e_{u,p}}$
```

Modeling Content Generation on Microblogs : Putting together the theme and event generation processes discussed above, TEM models each post p authored by user u as a mixture of a theme $z_{u,p}$ and an event $e_{u,p}$ (sampled as given by 1 and 2). The complete generative process is given in Table 1.

Scalable Inference: Our inference algorithm is motivated by collapsed gibbs sampling [3]. We employ forward sampling [1] which bases the estimates of the hidden variables on the data encountered so far. This provides us with a suitable process for carrying out the inference at the expense of suboptimal estimates at the beginning of the markov chain. This approach concurs with the online streaming of the data that is encountered on social media where data keeps pouring in and predictions should be made based on the data seen so far. In addition, we resort to a distributed implementation of the inference algorithm. All the operations are split across 7 threads with a master thread synchronizing all the operations. Data is read by the master thread 35K posts at a time and split equally across all the threads with 5K posts being handled by a single thread. Each thread maintains its own copy of the state that it modifies during the course of execution. 100 gibbs iterations are run over each batch of 5K posts in every thread. Thread states are synchronized upon completion of each such batch iteration.

3. 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 (300K posts obtained by extracting feeds from publicly available profiles over a span of three months) and Twitter (a subset of 20 million tweets crawled over a time span of about 2 months, [5]).

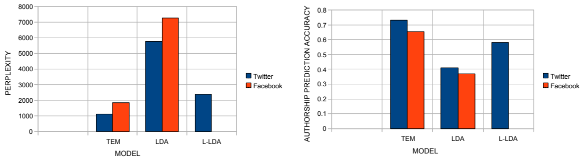

Figure 1: a. Perplexity of various approaches b. User Authorship Prediction Accuracy

Perplexity Evaluation : In order to compute model perplexity, we consider a sample of 3 million tweets spread out across last 15 day period and a sample of 40K posts from the last month's activity on Facebook. Figure 1a. depicts these measurements on Twitter and Facebook data for our model TEM and baselines Labeled-LDA [6] and LDA [2]. It can be seen that TEM gives the least perplexity in both the cases. Note that Labeled-LDA [6] cannot be applied for Facebook data as meta information like hashtags are not available on Facebook.

Predicting User Authorship : The question that we aim at answering is *Given a post p and user u, is user u likely to be the author of post p ?* We perform this task both on Facebook and Twitter datasets. In case of Twitter, a sample of about 3M tweets from the last 15 days of the crawled data is considered as the test set. For Facebook, we consider a sample of about 40K posts spanning the last one month as the test set. The results are presented in Figure 1b. It can be seen that our approach TEM performs significantly better on both Twitter and Facebook datasets. Labeled-LDA performs better than LDA, but in spite of using hash-tags, is significantly outperformed by our approach.

Additional Experimentation : TEM can be used for a wide variety of purposes like analyzing themes of interest for individual users and to capture events occuring at a particular point in time. We performed experimentation related to both these aspects. Due to space constraints, a detailed account of this experimentation is not provided here. For evaluating user interests, tweets from a set of 12 Twitter users was collected for a period of 15 days and then based upon their activity, the algorithm put forth a few themes that the users may be interested in. These users evaluated the themes output by the algorithm and about 76.32% of the themes suggested by the algorithm were interesting to the users. Further, while analyzing the events that our approach discovered, several significant events like demise of michael jackson, release of certain new movies etc. were well captured.

4. REFERENCES

[1] A. Ahmed, Y. Low, M. Aly, and V. Josifovski. Scalable distributed inference of dynamic user interests for behavioral targeting. In *ACM SIGKDD Conference on Knowledeg Discovery and Data Mining*, pages 373–382, 2011.

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

[3] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 2004.

[4] L. Hong and B. Davidson. Empirical study of topic modeling in twitter. In *KDD Workshop on Social Media Analytics*, 2010.

[5] J.Yang and J.Leskovec. Patterns of temporal variation in online media. In *ACM International Conference on Web Search and Data Mining*, pages 373–382, 2011.

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

[7] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 2006.