Modelling Retweet Dynamics using Hawkes Process - a temporal approach

Soumajit Pramanik, Agnivo Saha, Prithwish Mukherjee, Aseem Patni, Soham Dan, Bivas Mitra
Indian Institute of Technology, Kharagpur

Abstract

This paper presents a novel Retweet Model for online social networks.The objective is to model retweeting patterns based on historical data of user interactions, inherent topical similarity between tweets reaching current user and tweets of his top K friends, and the nature of interactions of a user with his neighbours in the social network. Point Processes have recently received significant attention from researchers in social media analytics. We have modelled Hawkes Process for online social networks to investigate retweeting patterns.

Introduction

Recently, there has been a massive surge in information content in online social networks,like, Youtube, Facebook and Twitter. Users share some information for all their friends or followers to view and reshare. The followers of a particular user view the information he has shared and may choose to reshare it depending on several factors. These factors are usually difficult to measure exactly but broadly include his/her topic interests, the extent to which he/she knows the user who has shared the information and whether other friends in that same circle are resharing that information. All these factors are crucial in influencing a user to reshare the information that is visible to him. Again, whether the information is visible to him depends on whether he is following the user who has shared the information or he has been mentioned in the tweet. All this leads to cascading of information. A fundamental problem in this domain is to measure the effect of this cascade and whether the shared information will be reshared by others or not.

Prediction of the information cascade is sometimes done by extracting a set of exhaustive features and then training a machine learning algorithm on it. A classifier trained on these features can predict whether a tweet will be retweeted or not. These are the two broad methodologies by which the problems in this area are addressed.

Stochastic Point Processes have recently found wide applicability by researchers in social media analytics.Point processes have been used for modeling in diverse areas of social media applications. They are typically used to model a phenomenon that occurs randomly over space or time. It has been used to model rumour dynamics in social media. (Lukasik, Cohn, and Bontcheva 2015) have used a log Gaussian Cox Process to model rumour dynamics. (Yang and Zha 2013) have modelled the spatio temporal aspects of meme spread by Hawkes Processes. Other authors have applied point process to other aspects of social media analysis like (Crane and Sornette 2008), (Ver Steeg and Galstyan 2012), (Zadeh and Sharda 2014) and (Zadeh and Sharda 2015).
Various models have been tried for the task of predicting retweet dynamics in the past. (Gao, Ma, and Chen 2015) have modelled the retweet dynamics using an extended reinforced poisson process model and they have predicted the future popularity of a tweet.(Zhao et al. 2015) have used a modified form of the Hawkes Process to model information cascades. Although, point processes have been explored by researchers to predict the size of the cascade, to the best of our knowledge, our work is the first that extends that to predict individual user's retweet behavior. Retweet behavior of a particular user depends on diverse factors that are often dependent on user interests, his neighbors and other factors and hence difficult to model. This is particularly crucial in order to recommend to a user, a set of users to mention given a tweet, based on the topic of the tweet, topical interests of users and other features. In this paper, we propose a retweet model that predicts the retweet behavior of individual users and is based on the Hawkes Process, a self-exciting point process.We have modified the generic Hawkes process equation to incorporate information of a user's past tweets, his topics of interest, influential neighbours' tweets and their interests and the structure of the network.

Methodology

Dataset

We have worked on a dataset which we have crawled. The dataset consists of tweets related to Arab Spring Movement in Algeria.The number of users is 19377 and the number of tweets including retweets is 54683. The number of tweets having retweets is 6453. The number of tweets having less than 15 retweets is 6140 and the distribution of tweets having more than 15 retweets is shown by Fig.

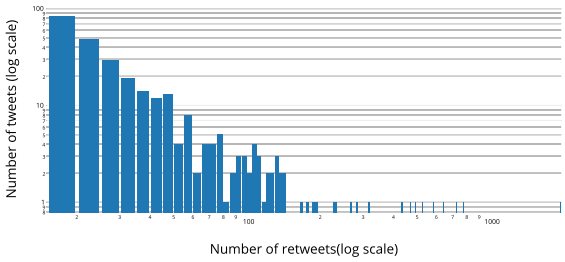

Figure 1: Dataset Retweet characteristics

Hawkes Process for Products

Hawkes process can be applied to learn the weights for predicting product adoption by a user. The equation (1) only considers influence of previous products by the user. So that the intensity of the product adoption at time t by a user becomes :

$$\lambda(t) = \lambda_0(t) + \sum_{t_i<t} \sum_{j=1}^{j=P} \alpha_j e^{-\beta(t-t_i)} \tag{1}$$

where $\lambda_0(t)$ is the bias for a particular user and β is the decay factor signifying the decrease in influence of a product the user may have used in distant past.

However, we are not considering the fact that the user may get influenced by his neighbours as well to adopt a product. Hence the Hawkes process (eqn. 1) for products needs to be modified to consider the influence that a neighbours may have on a user.

$$\lambda_p^u(t) = \lambda_0(t)_p^u + \sum_{t_i<t} \sum_{j=1}^{j=P} a_{jp}^u e^{-\beta(t-t_i)} + \sum_{t_i<t} \sum_{k \in Nbr(u)} \sum_{j=1}^{j=P} b_{jp}^u e^{-\beta(t-t_i)} \tag{2}$$

where $a_{jp}^u (b_{jp}^u)$ corresponds to the influence that a previous use of a product l by user u (by a neighbor of user u) has on user us intensity function associated to product p.

But, there are a few challenges for modelling the retweet prediction on Hawkes process. The total number of products i.e. the set of tweets in our case is very large. So, the assumption that the number of products is constant is invalid in our case. If we consider the tweets as products, then there is no upper bound on dimensionality of the feautes. As a result, we have used topics of the tweets instead of the number of tweets and modified the Hawkes Process eqn.2 further as described in the following section.

Modifying the Hawkes Process

We modeled the Hawkes process by representing the topic scores of a tweet which reaches a user as products and the phenomenon of buying the product is modeled as whether the user reshares/retweets the tweet upon receiving the tweet. The following is the modified Hawkes process equation that we have used in our work :

$$\lambda_i^u(t) = \alpha_u^u \cdot \sum_{j \in T(u)} TTS(ij) e^{-\beta(t-t_j)} + \sum_{k \in IU(u)} \alpha_k^u \cdot \sum_{j \in T(k)} sim(u,k) TTS(ij) e^{-\beta(t-t_j)} \tag{3}$$

where,
sim(u,k) - The influence of $User_k$ for $User_u$.
T(u) - Set of tweets tweeted by $User_u$ in the past.
TTS(ij) - Topical Similarity between $Tweet_i$ and $Tweet_j$, a $TX1$ vector where T is total number of topics considered.
IU(u) - Set of atmost 100 most influential neighbours of $User_u$
α_k^u is a $1XT$ vector representing the weights for each feature.
t - time when the current $Tweet_i$ is tweeted originally, t_j - time when $Tweet_j$ is tweeted
and $\lambda_i^u(t)$ predicts the probability that $User_u$ will retweet $Tweet_i$ at time t.

We frame the resharing problem as a classification problem. Each $(User_u, Tweet_i)$ corresponds to a training example and if $User_u$ has actually retweeted $Tweet_i$, we consider it as a positive example, else, we consider it as a negative example. For generating the TTS - Tweet Topic Score, we first normalize the tweets and use Latent Dirichlet Allocation to generate $TX1$ topic distributions corresponding to a tweet. We find the set of top K users for each user, that is, the set of Influential Users (IU($User_u$)) using the method described in next section. We treat the weights $alpha_u$, a $TX(N+1)$ matrix as constant for each user and learn the weight matrix using logistic regression, here N represents the number of neighbours of the user considered. The motivation to consider only top K neighbours instead of all neighbours is that we want to remove the bias which a user will possess due to having higher number of friends. We consider how likely is the $Tweet_i$ to be liked by $User_u$ based on the topical similarity with tweets which he has already tweeted. We also consider the topical similarity between $Tweet_i$ and the tweets which are reaching $User_u$ from his Influential neighbours, that is, the tweets tweeted by the Influential neighbours. There is a decay factor $e^{-\beta(t-t_j)}$, which ensures that the topics which the user tweets/sees recently to have a higher weightage than topics which he tweeted/saw in the past.

Finding the user-user influence We build a graph $G(V,E)$ where V = Set of all users in our training data
For 2 vertices, u_j, u_i, weight of edge (u_j, u_i) is defined as,

$$w(j \to i) = \begin{cases} \frac{|Retweet_i(j)| + P_j}{|TweetLink_i(j)|+1} & if(u_j, u_i) \in E \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where,

$$P_j = \frac{\sum_k |Retweet_k(j)|}{\sum_k |TweetLink_k(j)|} + \frac{1}{|Link(j)|} \tag{5}$$

$Retweet_i(j)$ - Number of tweets tweeted by $User_i$ that reached $User_j$ and were retweeted by $User_j$
$TweetLink_i(j)$ - Number of tweets tweeted by $User_i$ that reached $User_j$
$Link(j)$ - Number of users linked to $User_j$ We normalize the weights:

$$\forall i, j \; w^{norm}(i \to j) = \frac{1}{Z_i} w(i \to j) \tag{6}$$

where $Z_i = \sum_k w(i \to k)$
The motivation for choosing such weight was to model the influence of $User_j$ on $User_i$ by the ratio of tweets which the $User_i$ actually retweets among the tweets which he receives from $User_j$.
Now we have the graph $G(V, E)$. For each user $u \in V$, we do a Random Walk With Restarts on the graph G(V,E) and calculate the influence of the users accordingly. We then choose Top-K users, IU(u) for each user $u \in G(V, E)$.

Predicting if an user will retweet current tweet For each $User_u$ in our dataset, we found out the coefficients α_u^u and α_k^u as given by eqn.3 by using logistic regression on the set of tweets in $TweetLink(u)$ retweeted and not retweeted by $User_u$. Using the coefficients of the $User_u$, we calculate $\lambda_i^u(t)$ for $Tweet_i$ and $User_u$. We get the probability that $User_u$ will reshare $Tweet_i$ and based on the probability obtained, we predict if the user actually reshares/retweets the tweet.

Predicting users who will reshare the current tweet We consider the test set as mentioned in the next section. We perform a simulation of our model. For each tweet in the test set, we start the simulation with the friends of the user who has tweeted the tweet. For each friend of the user who has tweeted the tweet, we try to predict whether the friend will retweet the current tweet or not using our model. If our model predicts that the friend will retweet, then we add all friends of that friend in our set. If our model predicts that the friend will not retweet, then we ignore that friend. We continue this simulation until it converges, that is we have predicted for all possible users who have received the tweet according to our model, or we reach a maximum level of users in the user-friendship graph.

Evaluation

In our experiment, we select a test set of examples, (user, tweet) pair which were not used while training and try to predict whether the user will retweet the tweet. We report the precision, recall and accuracy values obtained. We also plot the ROC Characteristics curve and the Precision-Recall curve obtained by varying threshold used in the logistic regression.

Test set

We selected top 500 most retweeted set of tweets having atleast 60% of the retweets in the dataset. We use this tweet set to generate 2000 examples, (user, tweet) pair which we use for the evaluation, that is, finding accuracy of our prediction model.

Results

Retweet Prediction We obtained a Precision of 83.56%, Recall of 84.4% and Accuracy of 83.9% on the test data of 2000 examples as described in the previous section. The confusion matrix is given in Table . The ROC characteristics curve and the Precision-Recall curve obtained by changing the threshold of the logistic regression model at prediction time over the 2000 examples are given in Figures and respectively.

	Predicted true	Predicted False
Actual True	844	156
Actual False	166	834

Table 1: Test Set Confusion Matrix

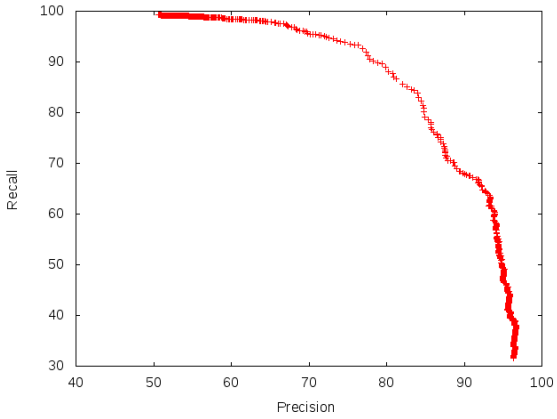

Figure 2: Test Set Precision-Recall Curve

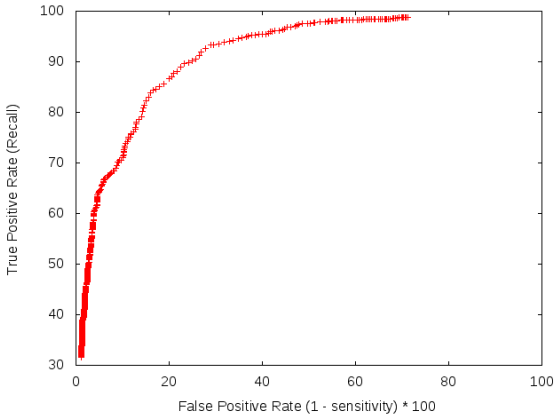

Figure 3: Test Set ROC Curve

Model Simulation Results We perform the simulation as explained in the Methodology section. We plot the mean and the standard deviations of the retweet count obtained over the set of 500 tweets. Level refers to the level in the breadth first search of the retweet network corresponding to

the tweet as obtained by our simulation. We plot the mean and median of retweets obtained by our model considering till level i where i ranges from 1 to 5. We also plot the mean and median of the 500 tweets as observed in the dataset, that is the true values. It is clear that with increase in level, the mean and standard deviation increases. As our dataset was extracted within a specific period, we feel that the tweets had not reached till level 5 in the actual retweet network, accounting for the anomaly obtained in the plot.

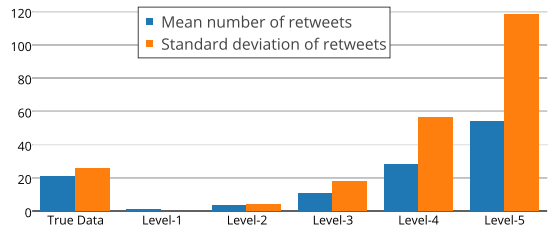

Figure 4: Test Set Mean and Standard deviation

Conclusion

Summary

We present a novel method of modelling the problem of Retweet prediction using the Hawkes process. This addresses two subproblems. One is to predict the retweeting probablity of the mentioned users or users whose friend has reshared the tweet, which in turn can be used to provide mention recommendations to the user posting the tweet based on the topics of the tweet. Secondly, by building the tree like structure for the friends of the mentioned users upto a certain depth and using the retweeting probability, we can predict the final number of reshares of the tweet, assuming no inactivity in the actual network. We have used topics rather than the actual tweets as our main feature reducing the complexity of the system. Also, using top K influential neighbours instead of all neighbours reduced the complexity of our system and removed the additional bias due to having a large number of friends.

Future Work

We are at present limited by the lack of disparate tweets in our evalution. As our primary future work we would like to extend our work to a much diverse dataset. We have only considered Hawkes Process. We would like to give a comparative study with other Point Processes like Poisson Process. We would like to consider other Machine Learning Models like Support Vector Machines and Neural Networks and find which models give better result.

References

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

Crane, R., and Sornette, D. 2008. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences* 105(41):15649–15653.

Farajtabar, M.; Wang, Y.; Rodriguez, M.; Li, S.; Zha, H.; and Song, L. 2015. Coevolve: A joint point process model for information diffusion and network co-evolution. In *Advances in Neural Information Processing Systems*, 1945–1953.

Gao, S.; Ma, J.; and Chen, Z. 2015. Modeling and predicting retweeting dynamics on microblogging platforms. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 107–116. ACM.

Hawkes, A. G. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58(1):83–90.

Lukasik, M.; Cohn, T.; and Bontcheva, K. 2015. Point process modelling of rumour dynamics in social media. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL*, 518–523.

Suh, B.; Hong, L.; Pirolli, P.; and Chi, E. H. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing (socialcom), 2010 ieee second international conference on*, 177–184. IEEE.

Ver Steeg, G., and Galstyan, A. 2012. Information transfer in social media. In *Proceedings of the 21st international conference on World Wide Web*, 509–518. ACM.

Yang, S.-H., and Zha, H. 2013. Mixture of mutually exciting processes for viral diffusion. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 1–9.

Zadeh, A. H., and Sharda, R. 2014. Modeling brand post popularity dynamics in online social networks. *Decision Support Systems* 65:59–68.

Zadeh, A. H., and Sharda, R. 2015. Hawkes point processes for social media analytics. In *Reshaping Society through Analytics, Collaboration, and Decision Support*. Springer. 51–66.

Zaman, T. R.; Herbrich, R.; Van Gael, J.; and Stern, D. 2010. Predicting information spreading in twitter. In *Workshop on computational social science and the wisdom of crowds, nips*, volume 104, 17599–601. Citeseer.

Zhao, Q.; Erdogdu, M. A.; He, H. Y.; Rajaraman, A.; and Leskovec, J. 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1513–1522. ACM.

Zhou, K.; Zha, H.; and Song, L. 2013. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, 641–649.