Mention Me

Prithwish Mukherjee (12CS10058)
Aseem Patni (12CS10008)
Agnivo Saha (12CS10062)
Soham Dan (12CS10059)

Social Computing Term Project
IIT Kharagpur

29-09-2015

Problem Statement and Motivation

Problem Statement

- Recommend mentions to a user posting a tweet in Twitter.

Motivation

- Twitter is one of the most important ways for information sharing.
- By mentioning users in a tweet, they will receive notifications and their possible retweets may help to initiate large cascade diffusion of the tweet.
- To enhance a tweet's diffusion by finding the right persons to mention, we propose a novel recommendation scheme named as "Mention-Me".

Introduction

Outline of work

- Modify Hawkes process to predict user-retweet probabilities.
- Obtain topic distribution of available tweets.
- Recommend users to mention in a given tweet.

Datasets

Datasets currenty used:

- Tweet outbreak related to 2010–12 Algerian protests.
  - Number of users - 19377
  - Number of tweets - 54683

Future datasets:

- Egypt Dataset. Currently under crawling,
  - Total number of users - 59776
  - Total number of tweets - 1350982

Motivations of using Hawkes Process

Point Processes like Hawkes Process are rapidly gaining popularity in modeling retweet propagation (or product adoption). Examples of related work being :-

- Trend detection in social networks using Hawkes processes, Julio Cesar Louzada Pinto, Tijani Chahed, Eitan Altman
- Modeling Adoption and Usage of Competing Products, Isabel Valera, Manuel Gomez-Rodriguez

Simple Point Process

Point Process

- In statistics and probability theory, a point process is a type of random process for which any one realisation consists of a set of isolated points either in time or geographical space, or in even more general spaces
- For example, the occurrence of lightning strikes might be considered as a point process in both time and geographical space if each is recorded according to its location in time and space.

Counting and Intensity process

Counting Process

Let (t_i) $i \in \mathbb{N}$ be a point process. Then

$$N(t) = \sum_{i \in N} \mathbf{1}_{t_i \leq t}$$

is called the counting process associated with (t_i) $i \in \mathbb{N}$.

Intensity Process

Let (t_i) $i \in \mathbb{N}$ be a point process. Then

$$\lambda(t) = \lim_{h \to 0} E\left(\frac{N(t+h) - N(t)}{h}\right)$$

is called the intensity process associated with (t_i) $i \in \mathbb{N}$.

Definition of Linear Self Exciting Process

Linear Self Exciting Process

A general definition for a linear self-exciting process N reads

$$\lambda(t) = \lambda_0(t) + \sum_{t_i < t} \nu(t - t_i)$$

where $\lambda_0 : \mathbb{R} \rightarrow \mathbb{R}^+$ is a deterministic base intensity and $\nu : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ expresses the positive influence of the past events t_i on the current value of the intensity process.

Equation

Simple Hawkes process

Hawkes(1971) proposes an exponential kernel
$\nu(t) = \sum_{j=1}^{j=P} \alpha_j e^{-\beta_j t} 1_{\mathbb{R}_+}$
So that the intensity of the model 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 P is the number of products.

Hawkes Process Application

- Apply Hawkes process to learn the weights for predicting product adoption by a user.
- The equation (1) only considers influence of previous products by the user.
- But user can also be influenced by his/her friends.
- Solution??

Modified Hawkes Process

Modified Equation

Consider influence of neighbours as well.

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

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 u's intensity function associated to product p.

Application for Mention Recommendation

Challenges involved:

- Number of products i.e. tweets in our case is very large.
- Number of products is assumed to be constant that is not valid in our case.
- Fix dimensionality of the tweet features.

Solution???

Equation

Solution

Fix the number of topics of the tweets instead of the number of tweets.

Modified Hawkes process

Fix the number of topics of the tweets instead of the number of tweets.

$$\lambda_i^u(t) = \alpha_u^u \cdot \sum_{j \in TweetLink(u)} TweetTopicSim(ij) e^{-\beta(t-t_j)} + \sum_{k \in Nbr(i)} \alpha_k^u \cdot \sum_{j \in TweetLink(k)} TweetTopicSim(ij) e^{-\beta(t-t_j)} \tag{3}$$

Equation

Terminologies:

- TweetLink(u) - set of tweets received by $User_u$ via a mention or a friend link or retweeted by $User_u$.
- TweetDist(i) = T dimensional Vector representing $tweet_i$ as probability distribution over T topics.
- TweetTopicSim(ij) - T dimensional Vector representing $tweet_i$ and $tweet_j$
  $TweetTopicSim(ij) = [TweetDist(k)_t * TweetDist(j)_t]_{TX1}$
- Nbr(i) - Friends of $User_i$.
- $User_i$ has (N+1)*T parameters to learn. (T = Number of Topics, N = |nbr(i)|)

Workflow

Preprocessing Dataset:

- Remove all tweets whose original tweet is absent.
- Remove all users with zero tweets or no retweets.

Preprocessing Tweets for better topical modelling:

- Remove all non-english tweets.
- Remove all stopwords.
- Replace terms by placeholders e.g. '$', 'pounds', 'dollar', 'dollar' by 'CURRENCY', '1000' by 'NUMERAL', '12:12 AM' by 'TIME' and so on.
- Replace 'Jan','January' etc. by 'MONTH'.
- Replace 'Mon', 'Monday' by 'DAY'.

Workflow

Topical Modelling

- Run LDA for 100 topics to obtain tweet topic distribution.

Regression

- For $User_u$, assign tweet ids in TweetLink(u) which $User_u$ has retweeted a value of 1 else value of 0.
- Learn the coefficients that is α_u^u and α_k^u for each user $User_u$ by Logistic Regression.

Workflow

Prediction

- Extract Features for $Tweet_i$, i.e. obtain TweetTopicSim(ij) $\forall j \in TweetLink(i)$ and obtain TweetTopicSim(il) $\forall l \in TweetLink(k) \forall k \in Nbr(i)$
- Use pre-obtained weight vectors α_u^u and α_k^u to classfiy the tweet.

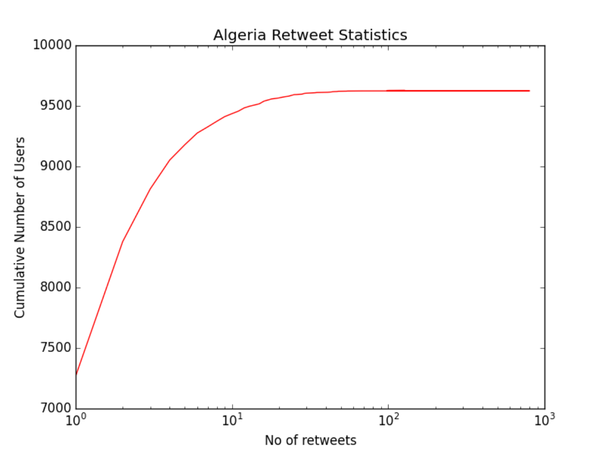
Evaluation

Evaluation strategy

- Sample a out random 450 tweets with Retweets and Non-Retweets in approximately 1:1 ratio.
- Train our classifier on remaining dataset.
- Predict whether a user will retweet a tweet in the test set.
- Compare it with truth values to obtain the confusion matrix.

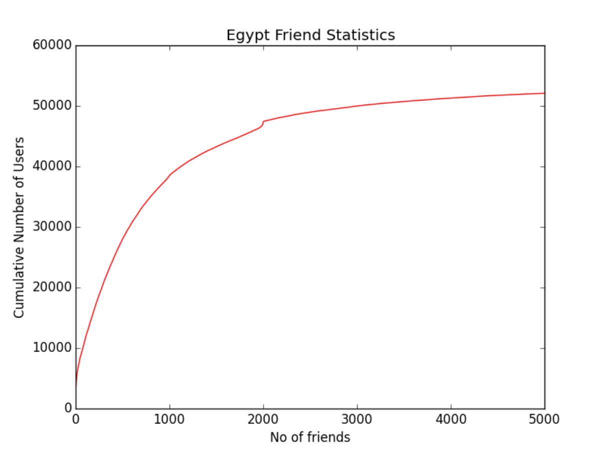
Some Dataset Studies

Algeria Dataset:



Some Dataset Studies

Egypt Dataset:



Next Step

- Train the recommendation engine for diversified datasets.
- Give a comparative analysis of various Point Processes (for eg. Poisson Process).
- Include similarity between users while considering feature contributions of neighbours.
- Proper evaluation strategies comparing to existing state of the art systems.
- Proper User Interface.
- Publish !!!

Post Mid-Semester

User - User Similarity

We build a graph $G(V, E)$.
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,

$$w(j \rightarrow i) = \begin{cases} \frac{|Retweet_i(j)|+P_j}{|TweetLink_i(j)|+1} & if(\mathrm{u}_j, u_i) \in E \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where,

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

$Retweet_i(j)$ - 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 $User_j$

User - User Similarity

We normalize the weights:

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

where $Z_i = \sum_k w(i \to k)$

Most Influential Users

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 $u \in G(V, E)$.

- This helps us to reduce the computation for training the model.

Modifying Hawkes' Process a bit more

Fix the number of topics of the tweets instead of the number of tweets.

$$\lambda_i^u(t) = \alpha_u^u \cdot \sum_{j \in TL(u)} TTS(ij) e^{-\beta(t-t_j)} + \sum_{k \in IU(i)} \alpha_k^u \cdot \sum_{j \in TL(k)} sim(u,k) TTS(ij) e^{-\beta(t-t_j)} \tag{7}$$

where,
sim(u,k) - The score of $User_k$ for $User_u$
TL(u) - TweetLink(u)
TTS(ij) - TweetTopicSim(ij)

Poisson Process

Now we simplify the Hawkes' process to get a much simpler model for recommending mentions.
Instead of considering the past retweets of an user as individual point processes we consider past retweets of different categories in a cumulative manner.

Poisson Process

$$
\begin{aligned}
\lambda_i^u(t) &= \alpha_u^u \cdot \sum_{j \in TL(u), topic \in T} TTS(ij) \frac{(\beta t)^{k_{topic}} e^{-\beta t}}{(k_{topic}!)} \\
&+ \sum_{n \in Nbr(i)} \alpha_n^u \cdot \sum_{j \in TL(u), topic \in T} TTS(ij) \frac{(\beta t)^{k_{topic}} e^{-\beta t}}{(k_{topic}!)}
\end{aligned}
$$

where
T - set of all topics
TL(u) - TweetLink(u)
TTS(ij) - TweetTopicSim(ij)
k_{topic} - Number of retweets of a topic by the user till time t

Evaluation Results

Confusion Matrix

.	Predicted True	Predicted False
Actual True	129	63
Actual False	19	225

Table: Hawkes' Process

Accuracy = 81.19%
Precision = 87.16%
Recall = 67.18%

Evaluation Results

Confusion Matrix

.	Predicted True	Predicted False
Actual True	135	57
Actual False	16	228

Table: Hawkes' Process with modification

Accuracy = 83.25%
Precision = 89.40%
Recall = 70.30%

Evaluation Results

Confusion Matrix

.	Predicted True	Predicted False
Actual True	114	78
Actual False	23	221

Table: Poisson Process

Accuracy = 76.83%
Precision = 83.21%
Recall = 59.37%

References

- Modeling Adoption and Usage of Competing Products, Isabel Valera, Manuel Gomez-Rodriguez.
- Wang, B., Wang, C., Bu, J., Chen, C., Zhang, W.V., Cai, D., He, X.: Whom to mention: expand the diffusion of tweets by@ recommendation on micro-blogging systems. International World Wide Web Conferences Steering Committee (2013)