Twitter Retweet Prediction

Debolina Mahapatra, 2019

Problem Statement

Since many years, Twitter has been an important platform in the study of social network analysis. Twitter supports "retweets", which is a means of information diffusion across the network. Given a tweet, we are required to predict whether it will be retweeted or not.

Review of existing systems

- Predicting users' re-tweeting behaviours based on the importance of content (eg. a trending topic) and interests of users
- Investigating the length of retweet chains of users
- Use of an epidemic model for the study of retweeting
- By studying individuals using their profiles, past retweeting behavior, their interests and the active times of the day
- Analyzing the retweet graph

Approach

- The Twitter social graph has been extensively studied in most of the works. However, the implicit retweet graph has not received much attention.

- In this work, the subtle differences between the social graph and the retweet graph have been analysed.

- The topology of the social graph has been considered for retweet likelihood in a number of literatures. In this work, we focus on the topological analysis of the retweet graph.

Dataset Description

The Higgs dataset (available on SNAP) has been used in this work.

The data was collected after monitoring the spreading processes on Twitter before, during and after the announcement of the discovery of a new particle with the features of the elusive Higgs boson on 4th July 2012.

The messages posted in Twitter about this discovery between 1st and 7th July 2012 are considered. The retweet network has been extracted from user activities.

The Retweet Graph

- In a RT graph, if A has re-tweeted B, then there is a directed edge from A to B.

- This graph is small-world, scale-free, less disassortative and has much stronger clustering. It more closely models real-world social and trust relationships among users.

- In our study, we have used the Retweet graph from the Higgs dataset. This network consists of 256,491 nodes and 328,132 edges.

Retweet Network statistics

The Retweet graph is in the form of a weighted edge-list, consisting of a source node, a destination node and the number of times the source re-tweets the destination (weight). It consists of 256,491 nodes and 328,132 edges.

Clustering coefficient :	**0.008**	Number of nodes :	**256491**
Connected components :	**13199**	Network density :	**0.0**
Network diameter :	**23**	Isolated nodes :	**0**
Network radius :	**1**	Number of self-loops :	**0**
Shortest paths :	**156883071 (0%)**	Multi-edge node pairs :	**758**
Characteristic path length :	**7.937**	Analysis time (sec) :	**1686.557**
Avg. number of neighbors :	**2.553**		

Topological Metrics

Taking the topological metrics into consideration, a dataset was constructed for the nodes involved in the retweet network structure.
The following are the topological metrics used as features:

- out-degree centrality
- in-degree centrality
- clustering coefficient
- closeness centrality
- authority centrality
- hub centrality
- eigenvector centrality
- pagerank centrality
- follower-followee ratio.

Modeling

- We model the retweet prediction problem as a binary classification task.
- Logistic Regression has been used as a GLM. It models data as outcomes of a coin flip and uses a Binomial Likelihood function, which perfectly fits our requirements.
- The topological parameter dataset was divided into 70% training data, 15% validation data, and 15% test data. 10 fold cross-validation was applied.
- The GLM based Logistic Regression model performs fairly for the binary classification task of retweet prediction and gives a precision of 98% and recall of 96%.

Issues

- The Retweet graph is a dynamic graph. Also, it is implicit and is derived from the social graph. A node involved in retweeting activity is only covered in the Retweet network. Other nodes in the social graph, which are not a part of the Retweet network (but may possibly become a part of it in future) are not considered.

- We may require additional features (eg: features derived from textual analysis). However, the dataset only includes anonymized social media profiles and their activity history and connections. We need to extract the respective tweets.

Future Work

- More features shall be incorporated and advanced learning techniques may be considered for modeling the Retweet Prediction problem.

- An extended dataset consisting of all the tweets along with the network data shall be used.

References

[1] Zhang et. al (2013), “Understanding Re-tweeting Behaviours in Social Networks”
[2] Remy et. al (2014), “Information Diffusion on Twitter: everyone has its chance, but all chances are not equal”, IEEE
[3] Goel et. al, “A note on modeling retweet cascades on Twitter”
[4] Bild et. al (2015), “Aggregate Characterization of user behavior in Twitter and analysis of the retweet graph”, ACM
[5] Domenico et. al (2013), “The Anatomy of a Scientific Rumour”, Nature Open Access, Scientific Reports 3, 2980