Complex Networks term project paper

Ivan Bogun

Computer Science department,
Florida Institute of Technology
ibogun2010@my.fit.edu

Abstract. Based on data downloaded from Twitter we present a study of complex networks. We use tweets about the movie "Captain Phillips" to build a word network. We show that the network is scale-free for each period of time. Using Naive Bayes classifier for sentiment analysis we predict movie rating as a fraction of tweets which were positive. We introduce the notion of community sentiments and show how they can be used to get a better insight of the network dynamics. We conclude the paper with experiments.

Keywords: Complex network, naive Bayes, twitter, sentiment analysis, scale-free network.

1 Introduction

Big Data era created an immense amount of information. Social media, being one of the largest sources of information on the web, is a great platform to mine information for commercial purposes. Product recommendations, opinions, polls all can be extracted from what people tweet all over the world. In the past Twitter was used to predict stock market prices [2], user influence [4], even the mood of the nation throughout the day [2].

In the current work we present a study which uses twitter data to build a network around tweets about a specific movie.

2 Related work

Due to wide commercial applications movie-related prediction gained much of an interest from the academia.

In [14] tweets and YouTube comments were used to predict movie ratings. Word frequencies were used as features for linear regression. In the current work we use tweet sentiments to predict ratings.

Work by [20] investigates whether it is reasonable to assume that tweets represent opinion from a general public. Their claim is that people who tweet are biased towards specific type of people and thus cannot be used to make reasonable predictions over the whole population. Authors introduce metrics
which show that tweets can be severely influenced by a "hype" which can be caused by outside events.

From tweet sentiment point of view pioneering work by [15] used different classifiers (naïve Bayes, SVM) to learn if a movie review is positive or negative. The problem later was extended to Twitter sentiments (see [10]).

3 Building a Twitter Network

3.1 Raw data

In this section we summarize when and how the data was downloaded and how the network was build.

The data was downloaded every day for six hours for a period of a month. Starting on October 12, 2013 (movie release day) and until November 8, 2013 everyday from 6:14 PM to 00:14 AM all English tweets containing the following keywords:

- CaptainPhillips
- Captain Phillips

were downloaded. The time was optimized so that the maximum amount of tweets could be downloaded. It was assumed that since most of people have to work during the day they will go to movies and tweet about them around the evening time. Table 1 shows distribution of the tweets downloaded per week. As expected the number of tweets is a decreasing function of time because the more time passed the more people who wanted to watch the movie did it.

<table>
<thead>
<tr>
<th>Week</th>
<th>Number of tweets</th>
<th>number of nodes</th>
<th>number of edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>October 12-18</td>
<td>4366</td>
<td>3253</td>
<td>36175</td>
</tr>
<tr>
<td>October 19-25</td>
<td>2251</td>
<td>2324</td>
<td>21480</td>
</tr>
<tr>
<td>October 26 - November 1</td>
<td>2280</td>
<td>2019</td>
<td>18035</td>
</tr>
<tr>
<td>November 2 - 8</td>
<td>846</td>
<td>1321</td>
<td>9835</td>
</tr>
</tbody>
</table>

Table 1. Time frame and amount of data downloaded.

3.2 Building a graph

Raw text data contains a lot of meaningless, incorrectly written words. In order to filter them out we for every word in the tweet we extract parts-of-the-speech (POS)\(^1\). Only words which were either adjectives, adverbs, nouns, verbs were left intact while the rest was deleted.

Once the data was preprocessed the graph was created in a following way:

\(^1\) we used POS parser from nltk library which was pre-trained on a large corpus. NLTK webpage: http://nltk.org/
1. for every word create a node
2. for every two words (nodes) appearing in the same tweet create an edge between them
3. set the weight of the edge as the number of occurrences of two words across all tweets

For further analysis we created such a graph for every week of the data. We also created a graph containing tweets for the whole month.

4 Experimental Results

In this section we evaluate graph characteristics of the networks. Firstly, we show that all four graphs follow power-log distribution. Further we provide numerical evidence suggesting that all four graphs satisfy “six degree of separation” [12] showing that all networks are scale-free. ADD other experiments

4.1 Degree distribution

Unlike random graphs many real complex networks follow power-log distribution of the node degrees [7, 3, 13]. It was natural to investigate if our movie networks follow the same degree distribution.

In [6] it was shown that plotting degree vs distribution in the log scale and calculating the scope of the trend can lead to inaccurate results. Authors suggested to use maximum likelihood estimate (MLE) instead. Assume that the degree follows Pareto distribution given by:

$$P(X \geq k) \sim k^{-\gamma}$$

where $X$ denotes a random variable of a node degree. Using maximum likelihood principle we can evaluate parameter $\gamma$. fig. 1 shows results while table 2 shows estimated values of the $\gamma$ for each week.

It was discovered that most scale-free networks have $\gamma \in (2, 3)$ [5]. Average $\gamma$ value across all networks was above 2 falling in the aforementioned interval.

<table>
<thead>
<tr>
<th>Week</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>1.94</td>
<td>2.05</td>
<td>2.05</td>
<td>2.21</td>
</tr>
</tbody>
</table>

Table 2. MSE estimate of the $\gamma$ parameter for each week.

4.2 Graph characteristics

It was natural to investigate if our networks are of the "small-world” class. According to the definition [1, 19] the network is called small-world if it has high
Fig. 1. Degree distribution versus degree in the log scale. Dotted line represents linear fit.

clustering coefficient and the distance between any randomly chosen two nodes in a network, denoted by \( L \), is given by

\[ L \sim \log N \]  

where \( N \) is the total amount of nodes. Small-world networks emerged after pioneering study by [12] who showed that in a social network average path between any two nodes is approximately 6, hence “six degree of separation” phenomenon.

Figure 2 shows how average path length and average clustering coefficient evolves in time. It should be noted that all four graphs can be characterized as small-world networks because all of them have high clustering coefficient and average path length satisfies eq. (2).

4.3 Sentiment analysis

There was a body of work where the problem was given aimed at problem of opinion orientation. Problem setup is as follows: having a text reviewing a subject find if the review is positive or negative [11, 18, 15]. Having huge commercial potential the problem was extended where tweets we used instead of reviews [9, 8]. Although tweet sentiment analysis is a harder problem because tweet consists
only from 140 characters current state-of-the-art method can achieve up to 85% accuracy [16].

In our analysis we opted to use basic sentiment analysis technique which is Naive Bayes model. In the model sentiment is a latent variable which is either positive or negative, each word is assumed to be independently of other words depend on the sentiment orientation (hence ‘naive’). It is easy to see that this is pretty strict assumption usually not satisfied, however the model was able to achieve reasonable accuracy in the sentiment analysis (see [17]). Formally let \( S \in \{\text{pos}, \text{neg}\} \) be a latent variable denoting if the tweet is positive or negative. Assume that the tweet has \( w_1, \ldots, w_n \) words. Naive Bayes assumption assumes that

\[
P(S|w_1, \ldots, w_n) \propto \prod_{i=1}^{n} P(S|w_i) \tag{3}
\]

holds true. Classification is done as maximum posterior probability:

\[
S^* = \arg \max_{s \in \{\text{pos}, \text{neg}\}} P(s|w_1, \ldots, w_n) \tag{4}
\]

We used nltk movie reviews corpus data for training. Results of classification on all tweets are shown in the table 3 where the first column denotes a fraction of tweets which were classified as positive, second as negative and the third column is a score of the movie from imdb.com\(^2\).

4.4 Sentiment analysis for communities

So far sentiment analysis was independent of the network structure. To get the best insight of the structure of the data we can calculate sentiments on the

\(^2\) Score from November 16, 2013
Table 3. Results of the tweet sentiment analysis

<table>
<thead>
<tr>
<th>positive</th>
<th>negative</th>
<th>IMBd score</th>
</tr>
</thead>
<tbody>
<tr>
<td>77 %</td>
<td>23 %</td>
<td>81 %</td>
</tr>
</tbody>
</table>

graph, like community sentiments. We define community sentiment as an average sentiment of the nodes in a community. Formally, let \( C \) be a community in the graph such that \( |C| = n \) it’s sentiment is given by:

\[
\mu(C) = \frac{1}{n} \sum_{w \in C} S(w) 
\]

(5)

where \( S \) is a function which maps tweet sentiments to a word sentiment. We define \( S \) as follows:

\[
S(w) = \frac{\# of positive tweets containing \ w}{\# of all tweets containing \ w}. 
\]

(6)

Figure 3 shows distribution for the community sentiments. Such metric can allow to identify what people like/dislike about the movie. Also we can query for a particular word and see how does it influence community sentiment in time.

\[\text{Only communities with higher than 20 elements are shown}\]
For example, if we query community which has the word ‘Tom’ we can track it in time and compare to average community sentiment. Figure 4 shows one example of a query. Results in fig. 4 suggest that people tend to appreciate Tom Hanks play more which might significantly correlate with his performance in the movie.

5 Conclusion

In this paper we studied word network build using Twitter text data about the movie "Captain Phillips". Firstly, we showed that network for each week is scale-free which can be explained by the fact that people tend to tweet similar to what was tweeted before (rich get richer). We also showed that high clustering coefficient and low average degree of the nodes in each network make them "small-world".

Using naive Bayes classifier for sentiment analysis we were able to classify tweets as being positive or negative. Results of classification were compared to the results from real movie ratings. In order to combine sentiment analysis with network structure we defined community sentiments and showed how they can be applied to query their progress in time.

References


4 Tom Hanks is the main actor in the movie

5 Week 4 was excluded because it produced one community with more than 95% of nodes in it