Temporal Analysis of Retweet Count in Twitter

Bavan Prashant, JunZe Han
Department of Computer Science, Illinois Institute of Technology, Chicago, IL
Email: bprashan@hawk.iit.edu, jhan20@hawk.iit.edu

***Abstract*—Twitter as an online social network offers an efficient way for information spread. Everyday a great number of tweets are posted every day and tweets exhibit rich and diverse temporal dynamics. Prior works have found that topics' popularities in Tiwtter *i.e.* mentions of hashtags exhibit some identical temporal patterns. However temporal patterns of individual tweet's popularity *i.e.* Retweet count and Favorite count have not been explored.**

In this work we study how the popularity of tweets grows over time, *i.e.* Retweet count and Favorite count increases. We analyze the time series of Retweet count and Favorite count and measure the similarity among their shapes to find certain common patterns. We then examine which factors of the tweets have impact on the patterns and whether can help predict the pattern that the Retweet count and Favorite count will follow. We analyze a set of 400 tweets from 20 users in one week. Our results show that based on the initial response to a tweet we are able to predict the temporal pattern with accuracy of 90%.

I. Introduction

In this research project we study the temporal patterns associated with a tweet's popularity *i.e.* Retweet count and Favorite cont, and what factors have impact on the popularity. Twitter offers an efficient way for information spread and a great number of tweets are posted every day. According to Statistic Brain [2] there are over 550 million active Twitter users as of May 2013, 58 million new tweets posted daily, and 135,000 new Twitter users every day. However some tweets get lots of Retweet in a short period of time after being post, while some receive only a little attention and gain a few Retweet s or Favorite s. As shown in Figure 1, among 450 tweets about 140 tweets got less than 100 Retweet s, while 10 of them are retweeted for over 1000 times. Accordingly, the temporal patterns by which the Retweet counts grow over time also vary among different tweets as shown in Figure 2.

In prior works, Yang [8] studied temporal patterns associated with online content and how the contents popularity grows and fades over time. They examined the number of mentions of Twitter hashtags over time and found there are several main temporal shapes of

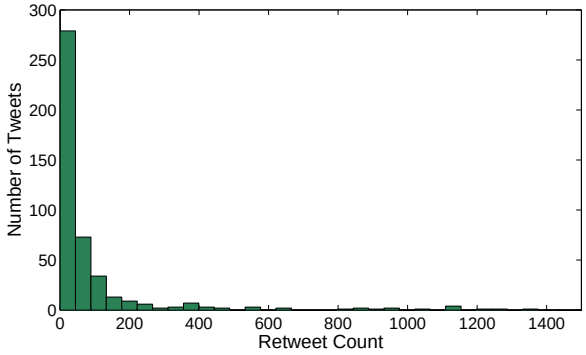

Fig. 1: Retweet Count

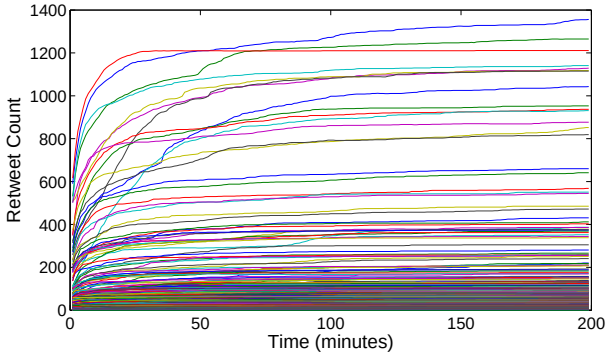

Fig. 2: Retweet Count

attention of online content. Asur *et al.* [1] studied what factors cause the formation and persistence of trending topics and found that the retweets by other users is more important in determining trends. Lehmann *et al.* [3] found that the evolution of hashtag popularity over time defines discrete classes of hashtags. However those previous works mainly focus on the analysis of the temporal pattern of topic spread *i.e.* hashtags which can be mentioned by a large number of tweets instead of individual tweets. In order to have a closer look at the

temporal variation of online content's popularity, we aim at uncovering their temporal patterns of individual tweets and factors that might have impact on their popularities.

In this work we study the temporal patterns of the popularity of individual tweets *i.e.* the Retweet count and Favorite count of a tweet. We analyze a set of 400 tweets from 20 users in one week. We would like to find good metrics to compare the similarity among tweets' popularities and then classify the temporal variations of the popularity into different categories. We formulate the problem of finding patterns of Retweet count and Favorite count as a time series clustering problem and measure the similarity among time series using Dynamic Time Warping (DTW) algorithm. Besides, we also study the correlation between the temporal patterns and the user's identification, the post time of a tweet and initial reactions to the tweet. Then given the patterns of the tweets *i.e.* clustering results of time series, we then examine effectiveness of different features for predicting the temporal patterns with different classification algorithms. Our results show that the initial resonance of a tweet in the first few minutes after the tweet is posted has a significant impact on the pattern of the Retweet count and Favorite count. In our evaluation, based on the Retweet count in the first 15 minutes we are able to predict a tweet's temporal pattern with accuracy of 90%.

II. Related Work

The temporal pattern of online contents has been extensively studied in prior works. Romero *et al.* [5] observed that different topics of hashtags in Twitter have different propagation pattern. They showed that the variation not only results from the 'stickiness', but also from 'persistence' — the extent to which repeated exposures to a hashtag continue to have significant marginal effects. Yang and Leskovec [8] analyzed a set of news articles, blog posts and Twitter hashtags and measure the attention given to various those contents by tracing the number of mentions over time. They developed a K-Spectral Centroid (K-SC) clustering algorithm to compute the cluster centroids and found that temporal variation of popularity of online social content can be accurately described by a small set of time series shapes. Asur *et al.* [1] studied trending topics on Twitter and found that the resonance of the content plays a major role in causing trends. Their results also showed that most topics break fairly quickly while there are few topics that last for long times. Lehmann and Goncalves [3] focused on peaks in the popularity of hashtags in Twitter. They found that the temporal pattern of hashtag popularity over time defines discrete classes of hashtags and epidemic spreading plays a minor role in hashtag popularity. Tsur *et al.* [7] presented a hybrid approach based on a linear regression for predicting the spread of an idea via social media. They evaluated their approach on Twitter hashtags and showed that content features can be used as strong predictors. However those works only studied the pattern of a topic, not the pattern of an individual piece of information.

III. Approach

A. Clustering

In order to discover identical temporal patterns among the time series of Retweet count and Favorite count, we perform clustering algorithm on the time series to group them into different categories. In our work we use K-means algorithm [4] to cluster the time series. K-means algorithm iterates a two step procedure, the assignment step and the refinement step. In the assignment step, K-means assigns each time series to the cluster closest to it. In the refinement step the cluster centroids are then updated. By repeating these two steps, K-means minimizes the sum of the squared distances between the members of the same cluster.

For clustering algorithm, instead of considering the time series as a vector and use Euclidean distance to compare the time series, we measure the shape similarity of two time series using dynamic time warping (DTW) algorithm [6]. In time series analysis, DTW is an algorithm for measuring similarity between two time series. More specifically, given two time series, and a cost metric, DTW finds an alignment that maps each point in the first series to one or more points in the second series, such that the sum of the cost of all mapping point pairs is minimized.

Let $t_1 = (c_1^1, c_2^1, ..., c_n^1)$ and $t_2 = (c_1^2, c_2^2, ..., c_m^2)$ be tweet t_1 and t_2's time series of Retweet or Favorite count respectively. Then we can compute the DTW distance between t_1 and t_2 as Algorithm 1.

B. Classification

Given the clustering results and patterns of a tweets' time series, we then predict what pattern the tweet will follow *i.e.* to which cluster the tweet will belong. We formulate the prediction problem as a classification problem. Let $(f_1, f_2, ..., f_m)$ be a set features of a tweet t and C_k be the cluster to which t belongs.

Algorithm 1: DTW

```
for i = 1 to n do
    DTW(i, 0) = ∞
for i := 1 to m do
    DTW(0, i) = ∞
DTW(0, 0) = 0 ;
for i = 1 to n do
    for j = 1 to m do
        DTW(i, j) = d(t_i^1, t_j^2) + min{DTW(i − 1, j), DTW(i, j − 1), DTW(i − 1, j − 1)}
    return DTW(n, m)
```

Then each tweet can be represented as an instance $< (f_1(t), f_2(t), ..., f_m(t)), C_k >$. We then perform two classification algorithms: Naive Bayes and Logistic Regression to predict the cluster of a tweet.

1) Naive Bayes: A naive Bayes classifier is a simple probabilistic classifier based on Bayes' theorem with strong independence assumptions. Let $P(C_k|t)$ be the probability that tweet t belongs to cluster C_k. According to Bayes' theorem, we have

$$P(C_k|t) = \frac{P(C_k)P(f_1(t), P(f_2(t)), ..., P(f_m(t))|C_k)}{P(f_1(t), P(f_2(t)), ..., P(f_m(t)))}$$

The basic assumption is that each feature f_i is conditionally independent of every other feature f_j for $j \neq i$ given the cluster C_k. Then we can estimate $P(C|t)$ as follows

$$P(C_k|t) = \frac{P(C_k)\prod_{i=1}^{m} P(f_i(t)|C_k)}{P(f_1(t), P(f_2(t)), ..., P(f_m(t)))}$$

2) Logistic Regression: Logistic regression measures the relationship between a categorical dependent variable and one or more independent continuous variables by using probability scores as the predicted values of the dependent variable. Let $g(C_k, t)$ be a linear predictor function for predicting the probability that tweet t belongs to cluster C_k, then we have

$$g(C_k, t) = \beta_{C_k}^T \mathbf{f}(t)$$

where $\boldsymbol{\beta}_{C_k}$ is is the set of regression coefficients associated with outcome C_k, and $\mathbf{f}(t) = (f_1(t), f_2(t), ..., f_m(t))$ is the set of feature associated with tweet t.

Then we can estimate the regression coefficients $\mathbf{f}(t)$ using maximum likelihood estimation.

IV. Evaluation

A. Data Collection

1) Social Network: Twitter, with its API, gives us direct statistics of a user and tweets at the current moment. In order to collect our data for the time series, we had to run through the list of available requests and get the most data without troubling the rate limit status that Twitter sets on us. So, we set a moderate data time capsule of 15 minutes, when we would refresh our data sets, for every user in our list. Although twitter has a reasonable API, they govern the usage with a rate limit to users. Hence, we define a well defined request set that we schedule to run every 15 minutes, to retrieve a time series of their retweet and favorite counts on every user's tweets.

2) Nodes: To make a good list of users with tweets with retweet and favorite counts to make up our list, we scout our users based on reasonably noticeable activity. To make the list neutral from biases of our selection, we select users from twitter's recommended users. We choose users, who are individuals and not organizations or groups. We select a mix of users, from three different categories. The complete list is described in Table I.

The users marked in Set1, are known to be quite engaging by retweeting and replying by tweets to followers. The users in Set 2 have a good following and see a lot of retweet and favorite activity on their tweets. The selection of these users was done to mimic a sample set of active users with quantifiable activity data. The line between the two sets is thin and we do not propose any validity in testing its contribution to the classification model we are trying to build.

3) Data: Moving on to the data collection, we set across the algorithms for collection to start on Oct 29th, and intended the run to last for a two week period. but due to some technical issues, the data, was not collected beyond 4th of November. We collected one week of data, that tracked 722 new tweets that appeared in the time frame. For all further classification problems, we decide to cut off all tweets with less than 250 count values collected every 15 minutes. Hence, the final total sample set of tweets we could work on, was reduced to 453.

Finally, we merge all data collected and create one dataset including the features shown in Table II. RT0, RT4, RT8 are retweet counts of the tweet at $<= 15$, 75 and 135 minutes and Fv0, Fv4, Fv8 represent the favorite counts of the tweet at $<= 15$, 75 and 135 minutes.

TABLE I: UserSet

Category	Set1	Set2
Fashion	ninagarcia, tyrabanks	heidiklum, victoriabeckham
Author	MargaretAtwood, SalmanRushdie	stephenfry, paulocoelho, DeepakChopra
Sports (F1)	JennieGow, NataliePinkham, JeremyClarkson	LewisHamilton
Technology	jasonfried, Scobleizer, jack, paulg	joshjames , fredwilson, marissamayer

TABLE II: FeatureSet

	data
UserDetail	UserID, UserName, Followers, Friends, isVerfied, <friend:followers>, ListCount, TotalTweet
Tweetdetal	tweetID, TweetDay, TweetHour
Rtdetail	RT0, RT4, RT8, RT198
fvdetail	Fv0, Fv4, Fv8, Fv198
<RtFvDetails>	<RTFv0>, <RTFv4>, <RTFv8>, <RTFv198>

RTfv stands for the cumulative of RT and FV at that specific time. The Cluster IDs are found on clustering the corresponding time series data over K-Means algorithms.

As you can see, the data set can be seen as three separate set of features. User details , Tweet Details and Time series data. Of these, there are some pieces of information such as ID numbers and text data, that are not usable in the classification process. They are still in the dataset for identification or back tracking of data.

B. Clustering

With all the data collected, we had to cluster them into groups for prediction afterwards. Ideally, with a given set of features, we should be able to predict the exact activity magnitude. But to make it a real world analysis, we start out by clustering the data set into groups and then predict the cluster they fall into. As shown in the feature set generated, we generate a cluster ID for the retweet count pattern, favorite count pattern and the combination of the two.We set out to divide the data into 5 clusters.

A sample representation of the color coded clustered data of Retweet counts is in Figure 3. It is a time series representation of retweet counts of tweets.

Some Statistical information on the clusteres identified for retweet count and Favorite counts are available in the Table III and Table IV

TABLE III: RetweetCluster

	C1	C2	C3	C4	C5
# of tweets	304	13	6	104	26
proportion	67.1%	2.8%	1.3%	22.9%	5.7%
avg # of retweets	17.01	939.69	1384.20	105.40	371.11

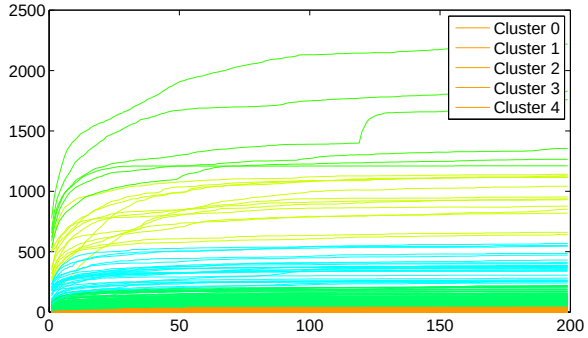

Fig. 3: Retweet Count

TABLE IV: Favorite Cluster

	C1	C2	C3	C4	C5
# of tweets	255	54	21	10	113
proportion	56.2%	11.9%	4.6%	2.2%	24.9%
avg # of favorites	3.56	164.46	386.57	924	53.41

C. Classfication

By varying the available features in the classification algorithm, we try to analyze the importance of every feature set. We run two algorithms to test and predict the cluster id from from the available features. A sample set of the features supplied and generated prediction success rates are listed in Table V. The table lists the input features used, the identified cluster, and the success rates using the Naive Bayes and Logistic Regression algorithms. All experiments were done by cross validation with 10 parts.

Our first goal was to identify the importance of feature

sets when they are held up with other features. We try a few pairs of feature sets and note the success rates in all casses.

TABLE V: Classification Results

Features	Class	NB	SL
UserDetail	RTCluster	66.22	66.44
UserDetail	FvCluster	59.161	60.26
UserDetail	RTFvCluster	60.48	60.76
UserDetail + RTDetails	RTCluster	89.1832	94.03
UserDetail + FvDetails	FVCluster	91.61	93.81
UserDetail + RTFvDetails	RtFvCluser	91.17	93.15
RTDetails	RTCluster	93.81	94.70
FVDetails	FvCluster	91.83	94.03
RTFvDetails	RTFvCluster	93.37	93.37
tweetHour + RTDetails	RTCluster	84.98	84.32
tweetHour + FvDetails	FvCluster	91.39	94.48
tweetHour + RTFvDetails	FvCluster	93.15	92.71
RTFvDetails	RTCluster	85.2097	84.10
RTFvDetails	FvCluster	79.0287	69.75
UserDetails + RTFvDetails	RTCluster	80.57	86.97
UserDetails + RTFvDetails	FvCluster	71.96	78.58

RTDetails, FVDetails and RTFvDetails include 3 Data counts of information available in their feature set. For example, RTDetails include RT0, RT4 and RT8. UserDetails include Followers, Friends, ¡friend:followers¿, ListCount, and TotalTweet. tweetHour is part of the tweetDetails we use as a single feature.

As you can see, the addiction of userDetails, doesn't seem to have any effect over the classification. Also, cross identification of features by supplying RTFV a cumulative value does not seem to help in identifying individual value.

V. Conclusion

The numbers from the tested data set seem to show very high success rates that do not seem realistic. A small data set, similarly active users, and time sheets with many identities matching make it a rather uninteresting classification problem. But these are just problems in the first run of the experiment, that helped show important pointers for future steps in the analysis. While there seem to be a many points of possible improvements, we were able to point to couple of interesting finds in the limited data set.

A. Results

1) Most items reached a saturation point at similar points of time.
2) The initial time series data was the most important feature in classification and prediction.
3) tweetHour had a hint of influence over Favorite cluster classification over using just the Favorite time series, which can be points to an interesting trend in usage.
4) UserDetails, did not seem helpful in differentiating activity clusters. We think this could be because of the similarity in the selected user's activity.
5) Tweet Detail had a better influence over the classification than userDetails
6) A combination of Retweet and Favorite count was not the as bad as using a non time series data to identify individual cluster ids.

B. Improvements

1) We feel that the data collected has some clues that a detailed extended analysis is possible on twitter data's time series. Even with such limited numbers, the option of identifying a user seem very much valid. To validate such work, we might need more data spread than just twenty users for less than a week. Re-running the existing program for the destined 2 week period will iron out this issue.
2) We felt the need to identify the retweeting user to calculate the magnitude of the specific users activity on the reach of the tweet. But this seemed slightly beyond the scope of this time period of research. But that definitely would give a larger graph data to identify feature sets.
3) The need to decrease the time period between the graph was felt when we tried reading the delta graph looking into the change per time period. There was a lot of activity in the first two hours, after which most curves seemed to saturate. Although the plan for this was underway, it was not put to test this time. A re-run of this experiment might help mine more closely monitored data.
4) User details did not seem to help the model in this data set. We feel that improving the sample set of users, and trying to get different visual patterns of activity would be a better sample set of the network. We do not have an explanation as to why user details have such an effect on the classification.
5) Identifying users, is not new, might already be in many advertising firm's technical know how. A leaf of understanding users to identify sample users and to test against their knowledge would give a good competing test to the model identified.

C. Division of Work

1) Data Collection: Bavan Prashant, Junze Han
2) Data Analysis: Junze Han, Bavan Prashant

REFERENCES

[1] ASUR, S., HUBERMAN, B. A., SZABO, G., AND WANG, C. Trends in social media: Persistence and decay. In *ICWSM* (2011).

[2] BRAIN, S. Twitter statistics. http://www.statisticbrain.com/twitter-statistics/.

[3] LEHMANN, J., GONÇALVES, B., RAMASCO, J. J., AND CATTUTO, C. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web* (2012), ACM, pp. 251–260.

[4] LIN, J., VLACHOS, M., KEOGH, E., AND GUNOPULOS, D. Iterative incremental clustering of time series. In *Advances in Database Technology-EDBT 2004*. Springer, 2004, pp. 106–122.

[5] ROMERO, D. M., MEEDER, B., AND KLEINBERG, J. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web* (2011), ACM, pp. 695–704.

[6] SALVADOR, S., AND CHAN, P. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis 11*, 5 (2007), 561–580.

[7] TSUR, O., AND RAPPOPORT, A. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining* (2012), ACM, pp. 643–652.

[8] YANG, J., AND LESKOVEC, J. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining* (2011), ACM, pp. 177–186.