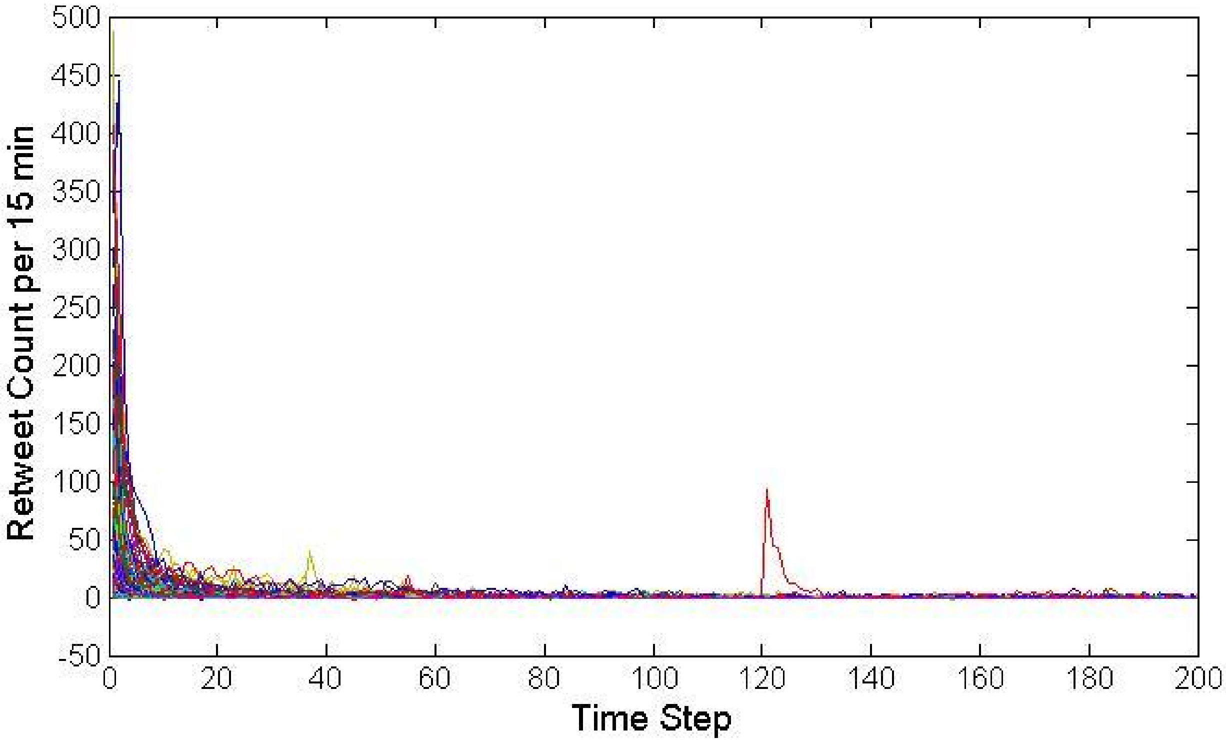
Temporal Analysis of Retweet Count in Twitter

By Bavan Prashant, Junze Han.
For CS 595 : Machine Learning and Social Network at IIT, Chicago. Fall 2013.

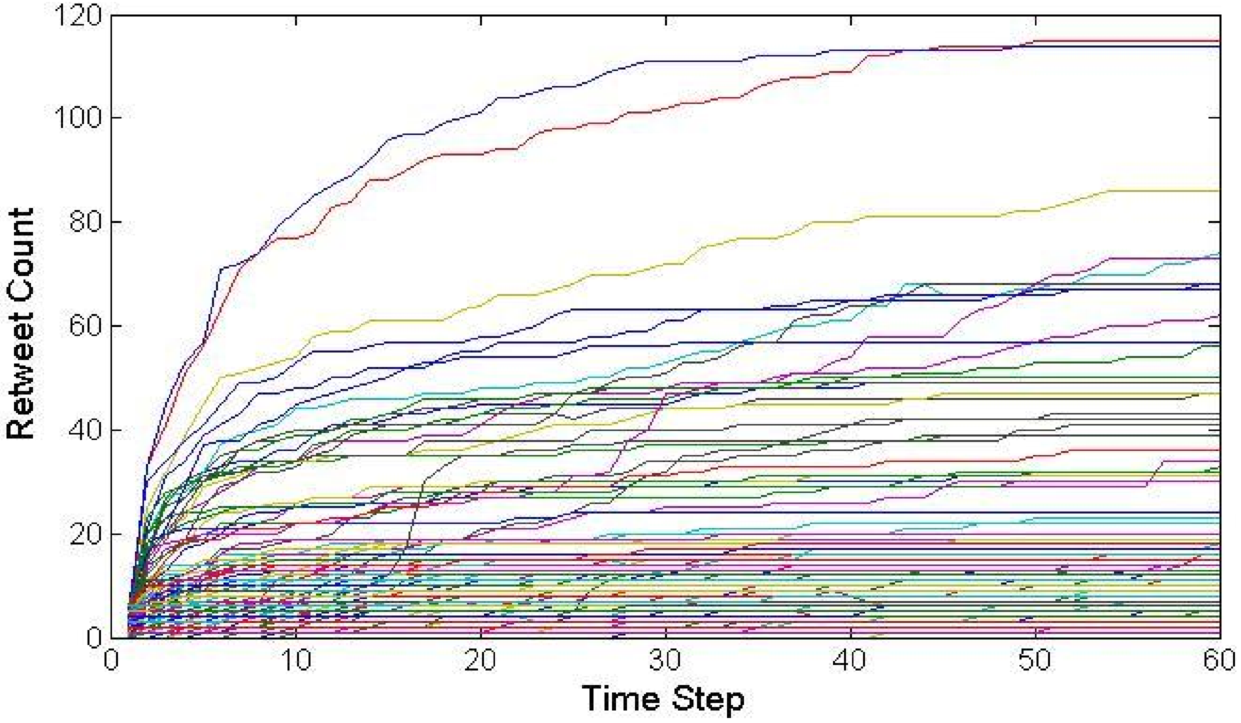
Motivation

- Analyze the time series of Retweets and Favorites counts in Twitter
    - How does the time series varies?
    - Whether there exist some patterns?
    - What factors have impact on the time series?
    - Can we predict what pattern a Tweet will follow?

Retweet Count: Per 15 Min

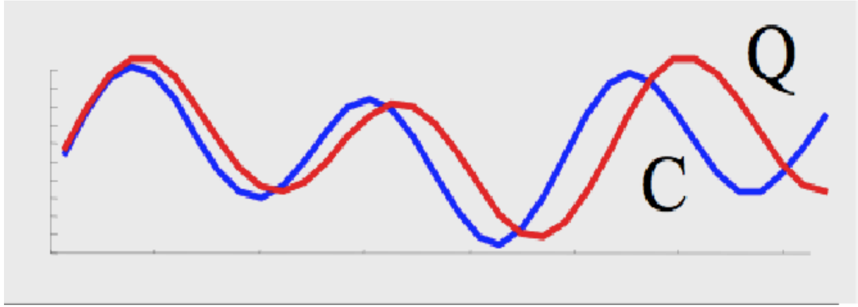


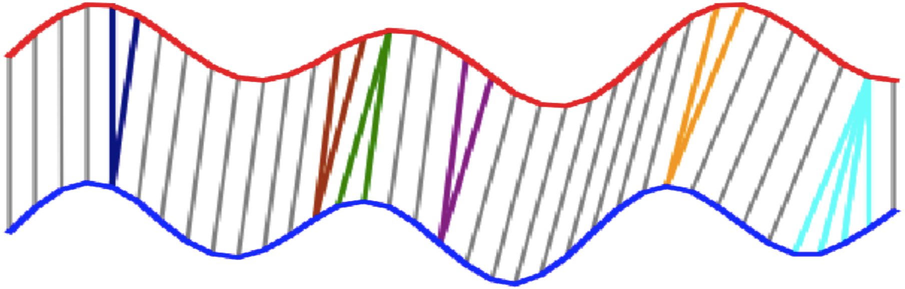
Retweet Counts: Cumulative



Group the Time Series

- Measure similarity between two time series
  - Dynamic time warping algorithm
  - dtw(i,j) =d(c_i,q_j)+ min{dtw(i-1,j-1),dtw(i-1,j),dtw(,j-1)}





Find Temporal Patterns by Clustering

- Finds clusters of time series that share a distinct temporal pattern
- Clustering time series of cummulative count into groups
- Clustering algorithm
    - DTW distance as distance metric

Predict the Temporal Pattern

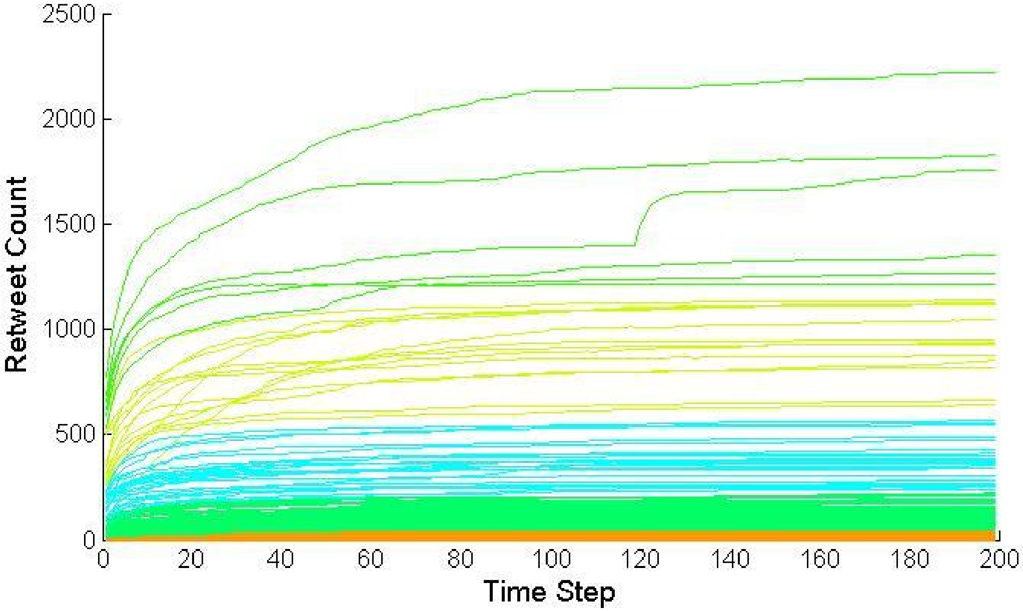
- Factors having impact on patterns
    - User attribute: # followers, # followers, ...
    - Tweet attribute: time and day
    - Initial activities: RTs and Fvs in first 15 mins
- Predicting the pattern as classification problem
    - Which attributes are good indicators
    - What classification algorithm to use

Data Collection

- Collect the status of users’ tweets over 2 weeks
    - New tweets every 15 minutes.
    - Retweet Count every 15 minutes.
    - Favorite Count every 15 minutes.
- 722 tweets in total
- 453 tweets with more than 250 timestamps
- We have 6 days* of continuous data from Oct 29th to Nov 4th, 2013.

Clustering

- K means algorithm
  - 5 clusters, 10 iterations, DTW as distance



Classification

- We group the attributes into 3 major parts.
    - User Details:
        - <UserID><UserName><isVerfied><Followers><Friends><friend: followers><ListCount><UserTotalTweetCount>
    - Activity Time Series
        - <RT0><RTFv4><RTFv8><RT198><Fv0><Fv4><Fv8><Fv198><RTFv0><RT4><RT8><RTFv198>
    - Tweet Details
        - <tweetID><TweetDay><TweetHour>
- Classify the cluster ID
    - RT Cluster ID, Fv Cluster ID, RTFv Cluster ID
- Classification Algorithms
    - Naive Bayes algorithm andLogistic Regression algorithm

Results :

Naive Bayes Algorithm

Cross validated split

[k=10]

User	RT Cluster	66.22	33.77
User	Fv Cluster	59.161	40.83
User	RTFV Cluster	60.48	39.51
User + RT	RT Cluster	89.18	10.81
User +Fv	Fv Cluster	91.61	8.38
User + RTFv	RTFv Cluster	91.17	8.83
RT	RT Cluster	93.81	6.18
FV	FV Cluster	91.83	8.17
RTFV	RTFv Cluster	93.37	6.625
RTFv	RT Cluster	85.20	14.79
RTFv	FV Cluster	79.02	20.97
User + RTFv	RT Cluster	80.57	19.42
User + RTFv	FV Cluster	71.96	28.03

Results :

Simple Logistic Algorithm

Cross validated split

[k=10]

User	RT Cluster	66.44	33.55
User	Fv Cluster	60.26	39.73
User	RTFV Cluster	60.75	39.29
User + RT	RT Cluster	94.03	5.96
User +Fv	Fv Cluster	93.81	6.18
User + RTFv	RTFv Cluster	93.15	6.84
RT	RT Cluster	94.70	5.29
FV	FV Cluster	94.03	5.96
RTFV	RTFv Cluster	93.37	6.625
RTFv	RT Cluster	84.10	15.89
RTFv	FV Cluster	69.75	30.24
User + RTFv	RT Cluster	86.97	13.02
User + RTFv	FV Cluster	78.96	21.41

Result/ Discussion.

1) User details tend to reduce effectiveness.

2) Is it possible to derive an individual cluster from derived values.

Example : RTFv Time series - RT / Fv

3) Will increase in time series help predict the cluster better ?

Pain Points / Discussions

- Our data collection failed !
    - Larger data would have helped critique the choice of algorithm.
- Smaller interval, might have had an impact?
- Day of week is not processed.
    - This is available though
- People who propagate! will it matter?
    - We identify the user being of of much help at this point! Can this get more proof?

References

1. Yang, Jaewon, and Jure Leskovec. "Modeling information diffusion in implicit networks." Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010.

2. Yang, Jaewon, and Jure Leskovec. "Patterns of temporal variation in online media." Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.

Thanks.