Association Rule Mining on Behavior Attributes extracted from Twitter Data

Chau, D. Thuan, SDSU Student, Lo, W. Donovan, SDSU Student

***Abstract*— Metadata is data that describes and contains information on another data. Metadata not only tell details about an instance of the other data, but it can represent trends when evaluated at a collective set. In this paper, we will collect and process the metadata extracted from twitter data and mine social media user 's behavior.**

In this paper, the first section will discuss about data collection and cleaning, followed by the methods that was applied to perform the behavior mining. Lastly, we will evaluated the effect that each parameters in the method have on the rules mined.

From this work, we hope provide new insights in the area of mining social media data by it's metadata.

***Keywords*— association rule learning, data mining, interestingness, lift, pattern detection, social media, social network analysis, Twitter**

I. INTRODUCTION

THE increase of social media usage has provided an abundance of data on both the data the media carries but also the metadata that help support the media. The data that the media carries contain interesting and meaningful metadata that help support the media. The data that the media carries contain interesting and meaningful data that the media carries contain interesting and meaningful information that the user wants to transport such as reports, images, accounts and etc. However, the metadata that is pertinent to that data only describe the instance of the data. It may contain fields such as created time, creator, attached entities' description, data format, and etc. The metadata alone only gives a snapshot on what, where, when the data was generated. By gathering a population of metadata for a particular group of users over a period, a record of the metadata can be utilized to show trends in the transactions.

The record of metadata collected will be accumulated in a table to summarize the frequency of each attribute. Each of the attribute will be quantized to labels that will be used for the Association Rule Mining.

The interesting rules will be selected from a list of ranked mined rules based on the interestingness and the confidence.

II. DATA PREPARATION

The Twitter data collected in this work uses the Julia Language and the Twitter API. The process consists of setting up the programming environment, Twitter account, gathering tweets, cleaning tweets, and structuring the data for association rule mining used in MATLAB.

In order to collect tweets from the Twitter API, a Twitter account and Twitter Application needs to be setup. From the created Twitter Application, the access token, access secret, consumer key, consumer secret can be extracted and used for the Twitter API.

To gather Tweets from two communities, San Diego and Los Angeles, we identified two Twitter account of origin relevant to the two locations (“KPBS”,”FOXLA”). From the two Twitter accounts, we gathered their followers as a list of Twitter IDs. To ensure that the tweets are from users of San Diego and Los Angeles, we verified each followers' account location and removed those that are not from San Diego or Los Angeles.

From the list of IDs in San Diego and Los Angeles, we do not know how much were active or non-active users, so we proceeded with Tweet collection. After three weeks worth of Tweet collection from followers from “KPBS” and “FOXLA”, we received 610,092 Tweets. Of all the Tweets, there were 5,603 active users from San Diego and 7,915 active users from Los Angeles. Active users is defined as have at least one Tweet. To make the samples from both locations to be balanced, we took the smaller size of the two groups as the sample size we use in Association Rule Mining.

For Association Rule Mining, we defined a set of metadata fields from the Tweets as the set of attributes that we like to do Rule Mining for. The metadata fields that we have considered are retweet count, user mentions, hashtags, favorite count, followers count, friends count, and created at. For each of the metadata fields, we quantized them into unique labels. For the field “created at” , we quantized the data “Thu Nov 17 23:56:12 +0000 2016” into few labels, weekend, weekday, work, off-work, and sleep hours. Work hours is defined from 8:00-17:00, off-work hours as 18:00-23:00 and sleep hours as 0:00-7:00. The other fields were binarized to either low or high with respective to designated threshold chosen by the median of that field.

After traversing through each metadata field in each of the collected Tweets, we tallied up each occurrence of the desired attributes. Due to the difference in units in the attribute label, we normalized the frequency of retweet count by per day, and hashtag, favorite, mention count by per tweet. The rest of the attribute is normalized by per week.

The final structure data from the Julia code that is fed to MATLAB for Association Rule Mining process is a csv file, where rows are users and columns are field labels.

The Julia Code is provided in the attached folder “Julia_Code” which contains a README file.

III. METHODS

The technique that we will use to extract interesting associated rule is the associated rule mining with pruning based on the Apriori Algorithm.

In the Association Rule Mining, we generated a frequent item set of attribute labels using a designated minimum support as the threshold to prune subtrees of item sets. The prune is made possible due to the Aprior Algorithm.

From the frequent itemset of potential candidate labels, we performed rule generation and reduced the list of possible associated rules by threshold of a designated minimum confidence of the rule.

A selective list of associated rule is then reduced by choosing the rules with the highest interestingness and confidence value.

IV. ANALYSIS

To extract interesting associated rule, we would like to select rules that have high confidence and high interestingness. However, before we choose rules of high confidence and high interestingness, let's look into the the effect of change in parameters Minimum Support, Minimum Confidence, and Minimum Interestingness.

By tuning the Minimum Support, we hope to see the number of generated rule to increase with a lower Minimum Support value and less rules with higher Minimum Support. By increasing the Minimum Confidence, we expect the generated rule to decrease. Lastly, if we increase the threshold of Interestingness, we expect a decrease in generated rule.

A. Change in Minimum Support for frequent itemset

Shared frequent itemset is the intersection between frequent itemset in SD and frequent itemset in LA.

Total frequent item set is the union of frequent itemset in SD and frequent itemset in LA.

Shared rules are intersection between association rules in SD and association rules in LA.

Total rules are union association rules in SD and association rules in LA.

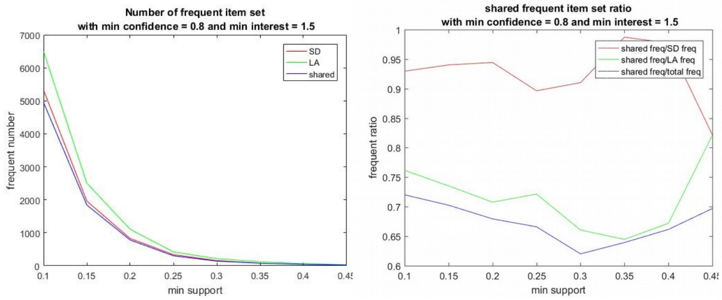


Figure 1. Right - Number of Frequent Itemset vs minimum support.
Left – Ratio for Frequent Itemset vs minimum support.

From Figure 1, we can see that as the minimum support increase, the number of frequent itemset for San Diego and Los Angeles decreases exponentially. Number of frequent itemset in SD for all cases are smaller than number of frequent itemset in LA. Moreover, the number of shared frequent itemset are very close to number of frequent itemset in SD. This can be seen in shared frequent ration figure, more than 80% of frequent itemset in SD are in the shared frequent itemset, as well as in frequent itemset in LA. The lines indicating shared frequent itemset/total and shared/LA are very close to each other, and line indicating shared/SD are above 0.8. In Figure 1,frequent item set is not affected by min confidence or interest.

B. Change in Minimum Support for association rules

From Figure 1, for association rules, increasing min confidence will lead to reducing number of rules significantly for SD and LA but there exists a pattern as we have for frequent itemset. As a given min confidence and min interest, when we increase min support, the number of rules for SD and LA decrease exponentially. Moreover, the number of rules in LA is much bigger than number of rules in LA (1.5 times bigger). Besides, the lines for number of rules in SD and number of shared rules are very close or similar. This shows most of the rules in SD are included in shared rules or association rules in LA. Without the extreme case which min support is 0.45, over 90% of rules in SD are shared rules as red lines in shared rules in Figure 1 is 0.9. In some cases, this red line goes up to 1. This is why the shape or value of shared rules/total and shared rules/LA are very close or similar.

One explanation for this pattern would be because most of frequent itemset in SD are in frequent itemset in LA. We could conclude that changing confidence does not affect this pattern based on the figure we have.

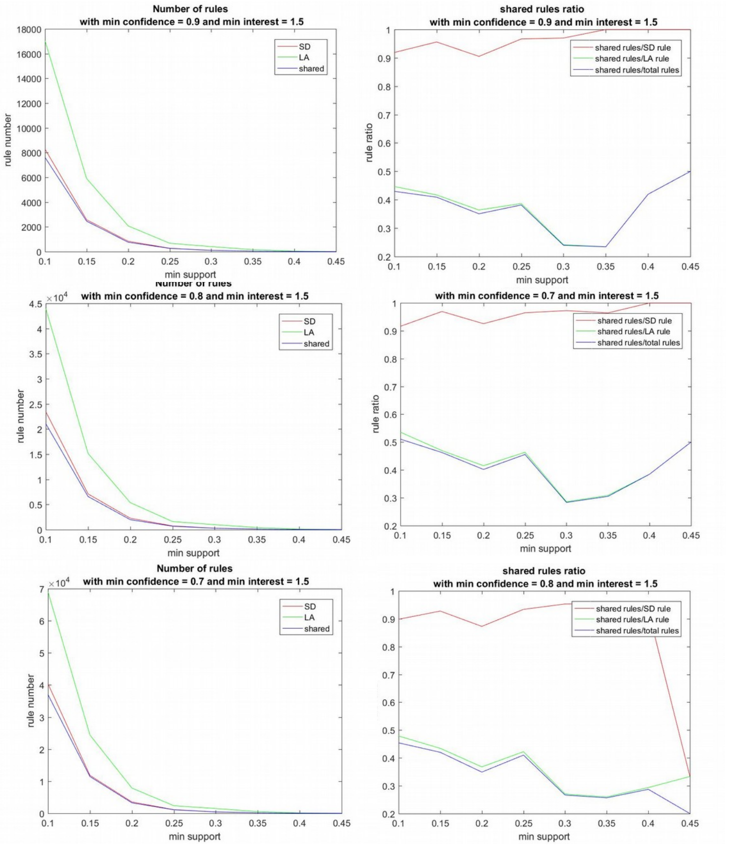


Figure 2. Right-side – Number of rules vs minimum support. Left-side Ratio of rules vs minimum support. Top-down is minimum confidence C=0.9, 0.8, 0.7

C. Change in Minimum Interestingness for number of rules

We set up min support is 0.4, as given min confidence, we try to change min interest to filter the number of rules we will have. (In Figure 3, min confidences corresponding to 0.6, 0.7, 0.8, 0.9). The number of rules reduce significantly when min interest is bigger than 1.4.

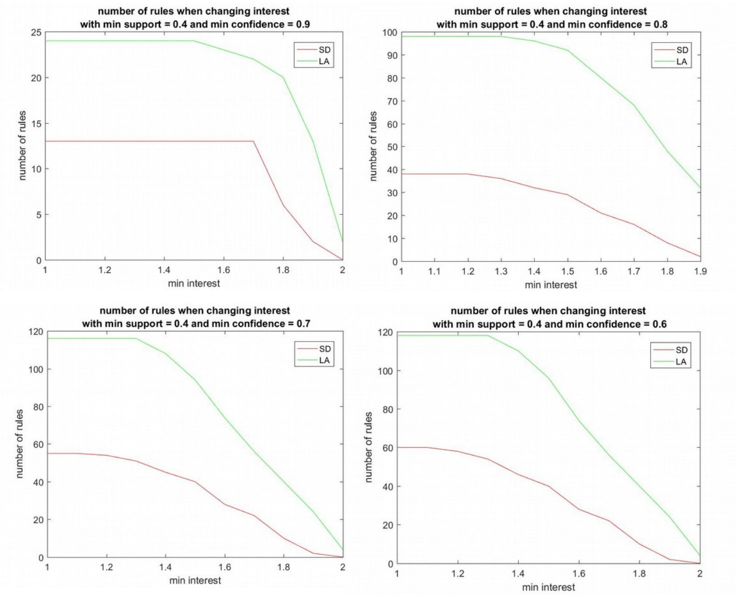
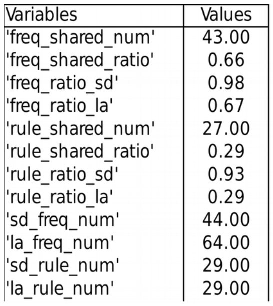


Figure 3. Number of rules vs minimum interest. Top-down, Left-right is minimum confidence C = 0.9, 0.8, 0.7, and 0.6.

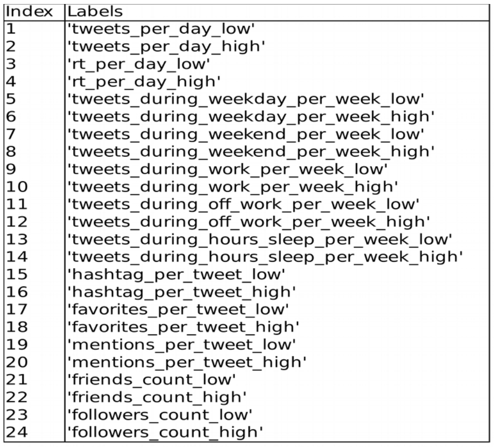
V. RESULTS

For the interesting associated rules we have collected, we used a minimum support of 0.4, minimum confidence of 0.8 and interestingness of 1.5. We were able to get 44 and 64 frequent itemsets for San Diego and Los Angeles respectively. Of those frequent itemset size, we received 29 rules for both locations.

From the the full list of associated rules that was gathered, we have ranked the rules first with Interest, and secondly with confidence. Of the associated rules with interestingness greater than 1.5, 10% of the rules in each location are exclusive and not shared.



Variables	Values
'freq_shared_num'	43.00
'freq_shared_ratio'	0.66
'freq_ratio_sd'	0.98
'freq_ratio_la'	0.67
'rule_shared_num'	27.00
'rule_shared_ratio'	0.29
'rule_ratio_sd'	0.93
'rule_ratio_la'	0.29
'sd_freq_num'	44.00
'la_freq_num'	64.00
'sd_rule_num'	29.00
'la_rule_num'	29.00



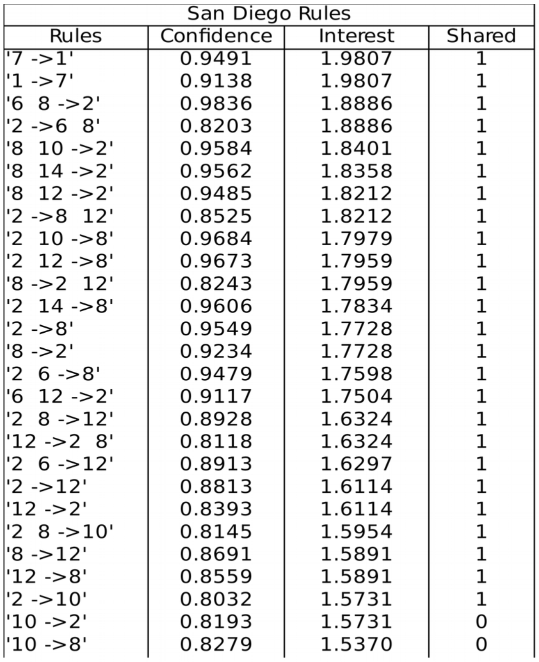
Index	Labels
1	'tweets_per_day_low'
2	'tweets_per_day_high'
3	'rt_per_day_low'
4	'rt_per_day_high'
5	'tweets_during_weekday_per_week_low'
6	'tweets_during_weekday_per_week_high'
7	'tweets_during_weekend_per_week_low'
8	'tweets_during_weekend_per_week_high'
9	'tweets_during_work_per_week_low'
10	'tweets_during_work_per_week_high'
11	'tweets_during_off_work_per_week_low'
12	'tweets_during_off_work_per_week_high'
13	'tweets_during_hours_sleep_per_week_low'
14	'tweets_during_hours_sleep_per_week_high'
15	'hashtag_per_tweet_low'
16	'hashtag_per_tweet_high'
17	'favorites_per_tweet_low'
18	'favorites_per_tweet_high'
19	'mentions_per_tweet_low'
20	'mentions_per_tweet_high'
21	'friends_count_low'
22	'friends_count_high'
23	'followers_count_low'
24	'followers_count_high'

Figure 4. Lookup table from Index to Labels.

Next we will evaluate the list of associated rules that were mined. Using Figure 4. , we can decode the index of the antecedent and consequent in Figure 5. to meaningful labels.

In the list of associated San Diego, we can see that the top ranked 25 rules are shared with Los Angeles. From the all the listed rules in San Diego, all the paired antecedent and consequent items are either all even or odd. From Figure 5, the even indice indicate high counts of a particular labels and odd indice indicate how counts of particular labels.

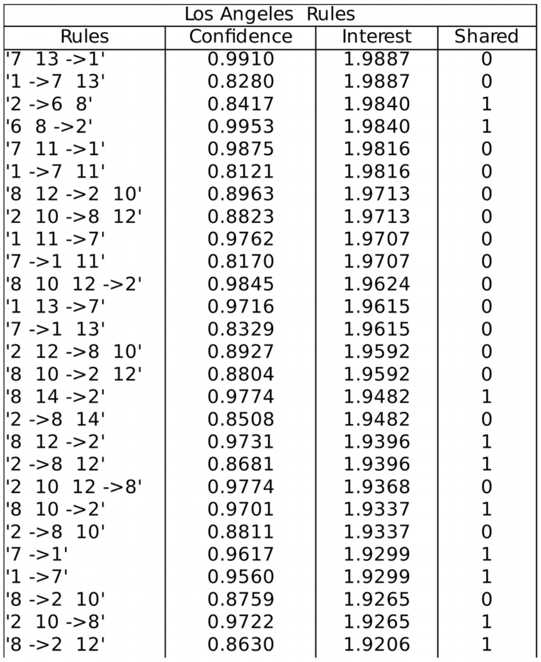
The top 8 associated rules are trivia, but the 9th rule may not be as trivia , Given that an active user that tweets a lot in a day and during work hours will tweet a lot on the weekend. Similar rule is depicted in rule 10, given user that tweets a lot in day and during work hours will tweet a lot on the weekend. However, rules 9 and 10 together formulates that a user that tweets a lot in a day not during sleep hours will be seen tweeting a lot on the weekend. Majority of San Diego's list of rules consist of either 1 or 2 as the antecedant or the consequent. This may not be as interesting since, it tells us that a user tweets a lot or less in a day. Let's examine the first rule that does not containing either 1 or 2, which will be rule 23 and 24. Rule 23 indicates that a user that tweets a lot on the weekend will tweet a lot during off work hours, which is defined from hours of 18:00-23:00.



San Diego Rules			
Rules	Confidence	Interest	Shared
'7 ->1'	0.9491	1.9807	1
'1 ->7'	0.9138	1.9807	1
'6 8 ->2'	0.9836	1.8886	1
'2 ->6 8'	0.8203	1.8886	1
'8 10 ->2'	0.9584	1.8401	1
'8 14 ->2'	0.9562	1.8358	1
'8 12 ->2'	0.9485	1.8212	1
'2 ->8 12'	0.8525	1.8212	1
'2 10 ->8'	0.9684	1.7979	1
'2 12 ->8'	0.9673	1.7959	1
'8 ->2 12'	0.8243	1.7959	1
'2 14 ->8'	0.9606	1.7834	1
'2 ->8'	0.9549	1.7728	1
'8 ->2'	0.9234	1.7728	1
'2 6 ->8'	0.9479	1.7598	1
'6 12 ->2'	0.9117	1.7504	1
'2 8 ->12'	0.8928	1.6324	1
'12 ->2 8'	0.8118	1.6324	1
'2 6 ->12'	0.8913	1.6297	1
'2 ->12'	0.8813	1.6114	1
'12 ->2'	0.8393	1.6114	1
'2 8 ->10'	0.8145	1.5954	1
'8 ->12'	0.8691	1.5891	1
'12 ->8'	0.8559	1.5891	1
'2 ->10'	0.8032	1.5731	1
'10 ->2'	0.8193	1.5731	0
'10 ->8'	0.8279	1.5370	0

Figure 5. Associated Rules ranked by Interestingness for San Diego.

Next let's examine the results in Los Angeles. From the list, a large portion of the rules are not shared with San Diego and are exclusive to Los Angeles. Similar to San Diego's list, the antecedent and consequence pair labels are either all even or all odd. The top rule in Los Angeles shows that users that tweet little on the weekend and tweet little during sleep hours will tweet little per day. The next associated rule that is high in interestingness and not shared with San Diego, is rule 5, which indicates that users who tweet little on the weekend and tweet little during off work hours will tweet little per day.



Los Angeles Rules			
Rules	Confidence	Interest	Shared
'7 13 ->1'	0.9910	1.9887	0
'1 ->7 13'	0.8280	1.9887	0
'2 ->6 8'	0.8417	1.9840	1
'6 8 ->2'	0.9953	1.9840	1
'7 11 ->1'	0.9875	1.9816	0
'1 ->7 11'	0.8121	1.9816	0
'8 12 ->2 10'	0.8963	1.9713	0
'2 10 ->8 12'	0.8823	1.9713	0
'1 11 ->7'	0.9762	1.9707	0
'7 ->1 11'	0.8170	1.9707	0
'8 10 12 ->2'	0.9845	1.9624	0
'1 13 ->7'	0.9716	1.9615	0
'7 ->1 13'	0.8329	1.9615	0
'2 12 ->8 10'	0.8927	1.9592	0
'8 10 ->2 12'	0.8804	1.9592	0
'8 14 ->2'	0.9774	1.9482	1
'2 ->8 14'	0.8508	1.9482	0
'8 12 ->2'	0.9731	1.9396	1
'2 ->8 12'	0.8681	1.9396	1
'2 10 12 ->8'	0.9774	1.9368	0
'8 10 ->2'	0.9701	1.9337	1
'2 ->8 10'	0.8811	1.9337	0
'7 ->1'	0.9617	1.9299	1
'1 ->7'	0.9560	1.9299	1
'8 ->2 10'	0.8759	1.9265	0
'2 10 ->8'	0.9722	1.9265	1
'8 ->2 12'	0.8630	1.9206	1

Figure 6. Associated Rules ranked by Interestingness for Los Angeles.

VI. CONCLUSION

In this project, we proposed a technique to predict Twitter user's social behavior by applying association rule mining on collection of metadata extracted from tweets. In the processed, we evaluated the effects of the change in minimum support and minimum confidence of frequent itemset and generated rules during pruning.

From the comparison we did on the associated rules on San Diego and Los Angeles, we identified that the top rules ranked by interestingness in San Diego largely shared with the rules in Los Angeles. However, Los Angeles top rules are not shared with San Diego.

Items that we could have improved on would be the twitter data clean up process. The associated rules that were generated may not contain as many interesting labels that we have expected. Some of the labels that we would like to see some of occurrences of in the top interestingness list would be favorites per tweet, mentions per tweet, friends count, or followers count.

If there was extra time to the project, we would have also like to extract rules of high interestingness and low confidence. By doing so , we hope to identify outliers, where the rules do not appear as often but has a high score of interestingness. These association rule may contain interesting patterns that is unexpected.

We hope that the work and findings in this paper to be insightful in the area of social medial user behavior prediction.

REFERENCES

] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In 4th International AAAI Conference on Weblogs and Social Media (ICWSM), 2010.

[2] M. Abbasi, and H. Liu, Measuring User Credibility in Social Media, poster paper, International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, April 2-5, 2013. Washington, D.C