Twitter Hashtag Prediction

Using Market Basket Analysis

Reilly Steele
Emily Cook
Jon Kennedy

DATA MINING TWITTER

Reilly Steel, Emily Cook, Jon Kennedy
Computer Science Department
Western Washington University
Bellingham, WA 98225, USA

Abstract **- Twitter is used for finding people and communicating with them across the Internet. Hashtags are used to group tweets by their topics to facilitate finding tweets about the same topic. We propose using association rule mining to be able to suggest hashtags that are relevant to inputted tweets.**

I. INTRODUCTION

For our project we decided to mine the social media site known as twitter. The goal of this project is to suggest hashtags to users when they are writing their tweets. This will be useful for twitter's mobile app since often users are trying to tweet on the go when they do not have much time. Having the option to see relevant hashtags rather than having to think about it would speed up their time, and typing on phone keyboards can often lead to mistakes. This will also be useful for users who struggle to know what relevant hashtags are, seeing as relevant hashtags increase the number of views your tweet receives. The more views a tweet receives the more likely it is to be retweeted, commented on, or even gain the user new followers.

II. BACKGROUND

The use of hashtags in a site such as twitter.com is to provide search functionality among groups of tweets. When a word contains a hashtag in front of it, it is transformed into a search-able word. Thus other twitter users can search for a specific word and they will get a list of all the tweets that contain that hashtag. For example if we had the following tweet "this is a tweet! #example," a Twitter user could click on "#example" and be given a list of all tweets that contain #example in them.

2.1 Definitions

Hashtag

A hashtag is represented by the pound sign (#). Hashtags are used through twitter as a way to search for tweets containing a specific word. For example, if somebody searched for 'baseball' they would get to see all the recent tweets containing the hashtag baseball. In addition for a way to search though tweet for a certain topic, hashtags are commonly used to gain more followers.

Tweet

140 character string that a user posts under their account.

Retweet

When a user re-posts another tweet to their own time-line.

Follower

A user that follows you. Which means every time you tweet it shows up in their news feed.

Support Limit

For our use, we're setting our minimum support at 5,000 tweets.

Confidence Limit

For our use, we're setting our minimum confidence at 50%.

Word Set

A word set for our project is defined as a single word or a grouping of words.

2.2 Software

Our data is taken from Twitter [1]. The Twitter stream gives a feed of a random sampling of new tweets written in English.

Our implementation was written in Python 2.7.

III. PROJECT

Our project consists of three parts – the dataset, the association rule implementation, and using the rules to predict the hashtags.

3.1 Dataset

We got our data from a json stream of recently posted tweets in the English language. We wrote this data to a text file and then preformed some data cleanup. A raw tweet contains a lot of unimportant information (the user who posted the tweet, the time, etc), so we use pythons json library to extract just the words and hashtags from the tweet. From there we continue to preform data cleanup by removing retweets and tweets that do not contain hashtags. From there we remove duplicated words from tweets so our support and confidence rules do not get skewed. Finally we remove conjunctions, prepositions, pronouns, and punctuation marks.

3.2 Association Rules Implementation

First, we build a Tweet object from the data, where a Tweet contains a list of words and hashtags. For example, the tweet "This is a tweet #example #python" would be broken into the following structure;

Words	Hashtags
this	#example
is	#python
a	
tweet	

Because existing aprioi algorithm code does not allow us to separate the antecedent of the rule from the original market basket, we implemented our own apriori algorithm:

First we calculate the frequency (in number of tweets) counts for all one item sets of words and delete the words whose support falls below our support limit. Then we use the remaining words to build all possible 2-word pairs, and calculate these pairs' frequencies, eliminating the pairs that now fall below our threshold. This process is repeated on 3, 4, …, n word sets where n is the largest word set length with support higher than our limit.

Second, we count the number of times each hashtag appears.

Finally, we calculate the confidence of each word set-hashtag pair by dividing the number of times that pair appears by the total number of times that hashtag appears.

3.3 Using The Rules

Once we get the association rules for our dataset, we can ask for tweets as input. We take the words inputted by the user of our software and suggest hashtags based on the words and the rules we have already found that are above a certain confidence threshold.

IV. RESULTS

We have produced words and word sets that relate to hashtags with certain levels of confidence. We are able to take new tweets and predict hashtags for them with high degrees of accuracy. For example, inputting the text “” into our software produces the following hashtag suggestions and their confidence levels:

V. CONCLUSIONS

In general, our research showcases that there is a relationship between words (and word sets) and hashtags while using twitter. We found out that this relationship is not symmetric, as often times a word A will have a strong confidence of hashtag B, but the word B does not have a strong confidence with the hashtag A.

VI. FUTURE WORK

Due to the time-sensitive nature of hashtags on Twitter, our data will need to be kept up to date in order to stay relevant. The issue is many hashtags revolve around events or are only used while a topic is “trending”, so suggesting an out-of-date hashtag will be useless. For example, we would not want to recommend the hashtag for last year's super bowl during next year's super bowl for users tweeting about football. In order to keep up with Twitter's rapidly changing environment, data collection would have to be kept running and our association rules would have to evolve with Twitter's conversations in real time.

REFERENCES

[1] Twitter Stream API (https://stream.twitter.com/1.1/statuses/sample.json?language=en)