Recommendation System based on Social Network of Brands

Abstract

Our main goal of this paper is to utilize the follower base of popular brands in United States on Twitter social network to identify the inherent attributes of the brands and recommend similar brands to consumers. Our approach measures the similarity of the brands based on their social network to find the most similar brand for a user based on the brands the user follows on Twitter. This is achieved by ranking and choosing top k brands which are not followed by the follower and predicting the brand which is very similar to the other brands followed the user using a baseline classifier developed and trained for this. Our initial findings on this approach has produced promising results for predicting the similar brands, which not only helps to recommend social media profiles to follow, but also to recommend products for similar potential consumers.

1. Introduction

Traditional recommendation systems for brands and products had used been primarily focused on utilizing the data about the purchasing behavior, product reviews, user's browsing history and social media content created by users. This paper explores and utilizes the structure of the social network of brands to recommend products to the consumers. Social Media presence of brands have proven to benefit them by not only develops the consumer relationship but also helps to improve the marketing strategy, to track the brand perception among the users and to improve the brand equity. Many consumers start following their favorite brands on the social media just to keep themselves updated with the brands most recent progress. Due to this increased presence and active follower network on the social media, analysis of the brands social network provides more information about the brands attributes and consumers preference about the brands. Our goal us to make use these social networks of brands to identify and recommend the brands of similar attributes to consumers.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Related Work

Various approaches to recommend a brand or a product have been primarily focused on the purchasing behavior of the consumers on e-commerce sites, reviews and opinions posted by the users expressing opinions. Many approaches were developed to understand the users' influence on social media by detecting the community structure (Cha et al., 2010) even by utilizing user-created lists (Bhattacharya et al., 2014). This had in turn helped to obtain further information like brand perception and brand attributes from the social network (Culotta & Cutler, 2016). Simulataneously several link prediction techniques were also developed (Liben-Nowell & Kleinberg, 2003; Hasan et al., 2006; Cukierski et al., 2011). These techniques and studies have guided us to attempt this novel method of predicting brands based on social media network.

3. Data

A list of Twitter handles of popular brands in United States were chosen across various sectors including Personal Care, Food & Beverages, Household Chemicals, Appliances, Apparel, Car, etc. For each of the accounts, the list of its follower ids were collected using Twitter API. Then each of the brands were manually categorized into different sectors. The list of all the sectors and the number of brands in each of the sector is given below:

Sector	Count
Apparel	24
Appliances	12
Baby Products	10
Car	18
Electronics	26
Food	226
Household Chemicals	33
Personal Care	205
Pet Food	11

3.1. Data Format

There were 1404 brands collected for this work. For each Twitter account, the list of Follower IDs were collected using Twitter API and stored in a tab separated format. Each row represents an adjacency list form of followers list for a brand. First column is the Twitter handle of the

brand and rest of the row is the list of follower ids with a maximum limit of 500000 followers. Only the Twitter handles of minimum 1000 followers were chosen for the experiment and any other brands of multiple sectors were eliminated from the experiment, leaving 565 brands and 35 million (35,462,028) unique followers for the analysis.

Sample Record:

```
cocacola      750103580686901248
2435842051       1546649604       3688890373
     1099956230       4707057674 ...
```

cocacola is the Twitter handle of *Coco Cola* brand and *750103580686901248, 2435842051,* ... are the followers (IDs) of the Coca Cola brand.

3.2. Initial Analysis

Before building the recommendation system, we performed various clustering analysis on the data set based on a similarity measure to find the natural clustering tendencies of the brands. Though we couldn't obtain a clear distinction of any brand attributes for each cluster, we noticed that in many clusters the brands and its competitors were getting grouped into one. For example, audi, bmw, cadillac and with few other brands were grouped into a single cluster. Since we have collected brands across multiple sectors and of unknown brand attributes, we were unable to assign a specific attribute to a single cluster based on the groupings. Perhaps with more diligent selection of brands and sectors of different known attributes could help us understand the clusters attributes better.

4. Methodology

Our methodology of building the recommendation system consists of the following:

1. Find the similarity between the brands followed by an arbitrary follower and all other brands not followed.
3. Score the brands not followed by the follower based on the similarity measure.
4. Choose top k brands as candidates to be fed as input for our baseline classifier.
5. Predict a brand using the trained baseline classifier.

4.1. Measure of Similarity

The problem of identifying the similarity of two groups of different size can be viewed as a central problem in many facets of social network analysis problems. Various studies (**???**) have also shown evidences that the Jaccard index addresses this problem better than many other measures. The Jaccard index $J(A, B)$ between two brands A and B can be calculated by the following relation:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where $|A \cap B|$ represent the number of follower overlaps between two brands and $|A \cup B|$ is the total number of unique followers of both brands A and B.

4.2. Scoring Brands

In order to find the brands that are similar to the ones followed by an user, first we need to score the set of brands $\overline{\mathrm{B_f}}$ that are not followed by a follower f. The brands $b_i \in \overline{\mathrm{B_f}}$ that are more similar to the set of brands B_f being followed by follower f must get higher scores, thus we calculate the Jaccard index measure of each of the brands $b_i \in \overline{\mathrm{B_f}}$ not followed by f with the brands $b_j \in B_f$ that are already being followed by user f and normalize based on the number of brands followed by the follower as below. For each $b_i \in \overline{\mathrm{B_f}}$, we calculate the score using the relation below

$$SCORE_{JS}(f, b_i) = \frac{\sum_{b_j \in B_f} J(b_i, b_j)}{|B_f|}$$

4.3. Baseline Classifier

After scoring the brands, we select the top k brands with higher scores for a follower and feed those brands as input to our baseline classifier. Our baseline classifier performs Logistic Regression on the inputs to provide a probabilities for each of k brands for a follower f. The brand with the highest rank (probability) is chosen as the most similar brand that the follower is likely to follow.

4.3.1. Features

(Liben-Nowell & Kleinberg, 2003; Hasan et al., 2006; Cukierski et al., 2011) provide various features that could be useful predict the missing links on social media network. In order to provide a brand as input to our baseline classifier, we represent each follower, brand combination (f, b_i) as a vector of various including:

- *Jaccard scores* $SCORE_{JS}(f, b_i)$ which we calculated in the previous step between b_i, b_j where $b_i \in \overline{\mathrm{B_f}}$ and $b_j \in B_f$,
- the *normalized number of brands followed the* f i.e. $\frac{|B_f|}{|B|}$ where $|B|$ is the total number of brands in our experiment,

- the *normalized number of followers for brand* b_i, i.e. $\frac{f_{b_i}}{|F|}$ where f_{b_i} is the set of followers who follow brand b_i and $|F|$ is the set of all followers,
- the *common neighbors score* $SCORE_{CN}(f, b_i)$ which is calculated similar to the Jaccard Score $SCORE_{JS}(f, b_i)$ by the following relation

$$SCORE_{CN}(f, b_i) = \sum_{b_j \in B_f} |b_i \cap b_j|$$

- *sector-wise brand count*, that is the number of brands in $b_j \in \{B_f \cap B_s\}$ where s is a sector from the set of all sectors S,
- *sector-wise similarity score*, that is the Jaccard index of brands between $b_i \in \overline{\mathrm{B_f}}$ and $b_j \in \{B_f \cap B_s\}$

In addition to the features mentioned, various statistics like mean, median, min, max values were also included as features making feature vector of dimension 41 with each feature scaled to 0 mean and normalized by its standard deviation.

4.3.2. Training

With the brands $b_j \in B_f$ that are being followed by a user f as positive samples and with brands $b_i \in \overline{\mathrm{B_f}}$ that are not being followed by the user f as negative samples, we train our baseline classifier for a set of arbitrarily chosen followers $f \in F$. Note that the feature vectors for the positive samples are generated after removing the link between the brand b_j and f however for the negative samples this is not necessary as $b_i \notin B_f$.

4.4. Performance Evaluation

In order to evaluate our classifier, we have separate set of brands which are chosen from a different set of followers (different from the ones used during training) whose links are removed similar to the ones in the positive samples during training and then the features vectors are generated. Following are the steps taken during evaluation phase:

1. Remove a random link between a brand b and its follower f.
2. Calculate the Jaccard index values between all the brands.
3. Calculate the Jaccard Scores $SCORE_{JS}(f, b)$ along with $SCORE_{JS}(f, b_i) \; \forall \; b_i \in \overline{\mathrm{B_f}}$.
4. Choose top k brands based on the Jaccard Scores as candidates.
5. Represent each of the candidates as a vector to be fed as input for the classifier.
6. Classify the candidates using the baseline classifier. The candidate with highest probability will be the most likely brand the user f will follow.

For a random set of followers different from the ones used during training is used in this phase and the above mentioned steps are performed. The accuracy of the system can be calculated for two cases by validating with the removed brand against the predicted brand. This is because, the final output depends on two phases, (a) elimination with Jaccard Scores and (b) prediction by our baseline classifier.

- Accuracy of the system when the removed brand was eliminated during (a) phase.
- Accuracy of the system when the removed brand is part of the candidate set i.e. not eliminated during (a) phase.

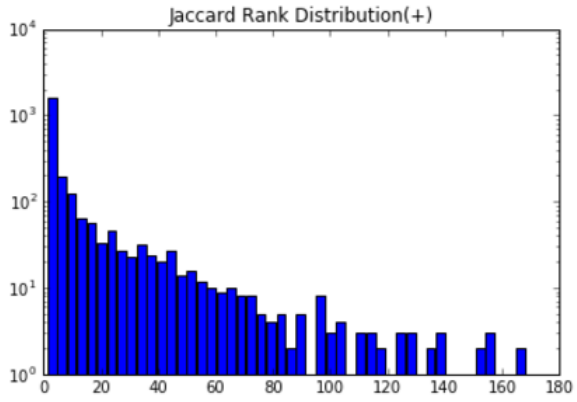
4.5. Results

In order to evaluate our system, we chose a random set of 200 brands from 565 brands for the experiment. Among the followers of these 200 brands who follow at least 5 brands, a set of 2434 followers were chosen for the training and 812 followers were chosen for the evaluation.

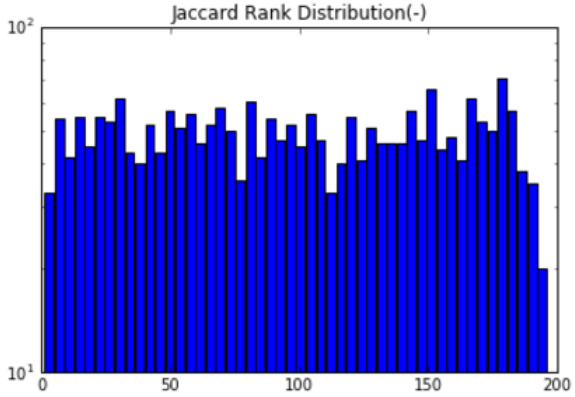
4.5.1. Baseline Classifier Training

Among the brands followed by each of the 2434 followers, a random brand was removed to be represented as a positive sample for the classifier. Thus 2434 positive samples were chosen for the classifier training. Similarly, among the brands not followed by each of the 2434 followers, a random brand was chosen as a negative sample for the classifier training, thus forming 2434 negative samples. Thus, for each of the 4868 samples, the jaccard scores $SCORE_{JS}(f, b_i)$ were calculated and the rank of each of these brands were analyzed before the training to validate our theory. We observed that for all the positive samples, the ranks of the removed brands were mostly higher unlike the ranks of negative samples which didn't have any specific pattern.

Rank distribution for the Brands in Positive Samples



Rank distribution for the Brands in Negative Samples



We could observe that the brands which were ranked lower among the positive samples had no similar brands that were being followed by the user. On the contrary, those brands which were ranked higher for the negative samples, were highly similar to some brands that were being followed by the user. These exceptional cases need more careful look so that the errors are minimized. Our logistic classifier was implemented in Python using the SGDClassifier class of sklearn's linear model module. This was chosen based on the GridSearchCV results. Based on a 10-fold cross validation training, the accuracy of the model was 0.861 with precision of 0.875 and recall 0.849. Out of all the features, Jaccard Scores $SCORE_{JS}(f, b_i)$, common neighbor scores along with the statistics were found most informative whereas some of the sector features were the least as we could see that the number of brands per sector was skewed with most of the brands in only two of the sectors personal care and food & beverages.

4.5.2. Evaluation

A set of 812 followers were randomly chosen and a brand from each of the followers is removed. All brands not followed by each of the followers were scored and top 25 of the brands based on the Jaccard scores were chosen as input for the classifier. This is because we noticed that the average rank of the positive samples were ranging between 11 and 23. Thus, we chose the value of k as 25 for our experiment for choosing the candidates. We noticed that about 143 of the 812 brands were not appearing in its respective candidate sets. The classifier was able to predict 49.02 % of the brands correctly without considering the brands that did not appear in the candidate set. However, if we consider the overall accuracy of our recommendation system, the accuracy was 40.4 %. By choosing a higher value for k, we can ensure most of the brands appear in the candidate set, however the accuracy of the classifier in this case drops further to approx 34.2 % for $k = 50$. This trade-off between choosing k and the classifier's accuracy need to be studied further to tune value of k and the parameters of the classifier so that the overall system's performance is improved.

5. Conclusion and Future Work

Though the Jaccard Scoring method provides a decent measure to start with this system, more research is needed here to improve the scoring method so that the exceptional cases as mentioned above are handled by a scoring method better that a simple average of the Jaccard scores. 143 out of 812 brands is a substantial amount for the classifier to make errors. Thus, a better weighted scoring method similar to the SPS scoring method in (**?**) along with the purchasing behaviour of the consumers could improve the ranks of the exceptional cases seen during candidate selection phase for the better. Though we have attempted to include the sector information tot he feature vector, we could consider adding additional sector specific attributes to the features vector to improve the accuracy of the classifier. Also, identification and removal of the bot from the followers our data should also improve the accuracy of our system.

References

Bhattacharya, Parantapa, Zafar, Muhammad Bilal, Ganguly, Niloy, Ghosh, Saptarshi, and Gummadi, Krishna P. Inferring user interests in the twitter social network. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pp. 357–360, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2668-1. doi: 10.1145/2645710.2645765. URL http://doi.acm.org/10.1145/2645710.2645765.

Cha, Meeyoung, Haddadi, Hamed, Benevenuto, Fabrcio, and Gummadi, Krishna P. Measuring user influence in twitter: The million follower fallacy. In *in ICWSM 10: Proceedings of international AAAI Conference on Weblogs and Social*, 2010.

Cukierski, W., Hamner, B., and Yang, B. Graph-based features for supervised link prediction. In *The 2011 International Joint Conference on Neural Networks*, pp. 1237–1244, July 2011. doi: 10.1109/IJCNN.2011.6033365.

Culotta, Aron and Cutler, Jennifer. Mining brand perceptions from twitter social networks. *Marketing Science*, 35(3):343–362, 2016. doi: 10.1287/mksc.2015.0968. URL http://dx.doi.org/10.1287/mksc.2015.0968.

Hasan, Mohammad Al, Chaoji, Vineet, Salem, Saeed, and Zaki, Mohammed. Link prediction using supervised learning. In *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*, 2006.

Liben-Nowell, David and Kleinberg, Jon. The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pp. 556–559, New York, NY, USA, 2003. ACM. ISBN 1-58113-723-0. doi: 10.1145/956863.956972. URL http://doi.acm.org/10.1145/956863.956972.