Graph based User Interest Modeling in Twitter
CS 224W : Final Project , Group 41

Aparna Krishnan
aparnak@stanford.edu

Raghav Ramesh
raghavr@stanford.edu

1 Motivation

Interest graph is a comparatively recent phenomenon in social media, building on the lines of Knowledge Graph[1] and Social Graph[2]. Simply put, an interest graph is a representation of relationship between people and things that they are interested in. Companies like Netflix and Facebook are already leveraging Interest graph to power their recommendations. Additionally, these interest graphs fuel behavioral profiling based on interests for advertisement targeting and audience analytics. One social network that inherently presents itself as an interest graph is Twitter. In this project, we explore a graph based approach to user-interest modeling in Twitter.

2 Related Work

In the light of the current push towards development of an interest graph, identification and characterization of user interest in social media is very relevant. Most works focusing on identifying user interest on Twitter use text analysis techniques like Bag of Words, TF-IDF or Latent Dirichlet Allocation [Michelson and Macskassy (2010)]. This paper proposes a method to discover twitter users topics of interests based on the entities in their tweets. Another common method used to extract user interests is to use a knowledge corpus as in [Lim and Datta (2013)] where the authors infer the users topics of interest from the domains of the celebrities followed by the user. Taking cue from these works, we hypothesize that since Twitter as a social medium is primarily based on who we follow, the properties of network structure of a user are potentially indicative of his/her interest profile.

3 Problem Definition

The high level goal of the project is to study the following two questions:

1. How well can we gauge the interest profile of a user based on his network structure?
2. How predictive are different graph based features in identifying the interests of a user?

The intuition behind this is the following: a user interested in baseball will follow his favorite team, say @SFGiants, while one interested in technology will follow @TechCrunch. Thus, the network of a user represents his/her interest. Additionally, the more a user (say, a huge fan of Star Trek) is interested in a topic, higher will be his interaction with @StarTrekMovie.
This, the primary goal of this project is to model the interest profile of a Twitter user using the network structure of the Twitter follower-friend graph. Concretely, we explore various graph features to capture the level of influence exerted by the followers and friends, and use this to predict the interests of the user. Additionally, we use a machine learning approach to build a supervised model to predict the interest profile of a user.

4 Data

4.1 Data Collection

We obtained Twitter social graph data from the internet[3]. The dataset contains 41.7 million nodes and 1.47 billion edges. A dataset of tweets was obtained from SNAP datasets. This data contains 476 million tweets that includes about 50 million hashtags and 180 million URL entries[4]. Since the social graph dataset references users by a numeric user-id and the tweets datasets does it through the users screen name, a user-id to screen name map obtained from [3] was used to link the two datasets.

[1] www.google.com/insidesearch/features/search/knowledge.html
[2] www.facebook.com/about/graphsearch
[3] http://an.kaist.ac.kr/traces/WWW2010.html
[4] http://snap.stanford.edu/data/bigdata/twitter7/

4.2 Data pre-processing

Since, we use data from different sources, a considerable amount of pre-processing was required to get the data to a usable form. We use a MongoDB to store all data required for the project. We start with the tweets dataset and use the user-id : name map to insert the data into a db collection with the following schema - user_id, name, [list of tweets] . From this, we subset users who have a minimum of 50 tweets. Then, we use the Twitter graph data to obtain the list of followers and followees for these users and form our dataset.

5 Mathematical Model

In this section, we describe the mathematical formulations of our project and define the terms that are used in the rest of the report.

5.1 Identifying relevant subgraphs

As discussed in the data description, the Twitter graph we deal with contains 41.7 million nodes and 1.47 billion edges. Such a large graph can clearly not be dealt with in its entirety due to the memory and processing limits. The natural alternative is to consider only neighborhoods that relate to the node in focus. In fact, this fits our bill perfectly as only nodes that are closest to the node in focus, can possibly influence its interest profile. Let G_a denote the subgraph for a node *a* , containing all nodes and edges that influence *a*. We used two methods of arriving at the subgraph, denoted G1 and G2,

- $G_1(a) = \{\Phi(a) \cup \Psi(a)\}$
- $G_2(a) = \{G_1(a) \cup \Psi(G_1(a)) \cup \Phi(G_1(a))\}$

where

- $\Phi(a)$ denotes the set of followers of *a*
- $\Psi(a)$ denotes the set of friends (those who follow *a*) of *a*.

The subgraph G_1 is useful in quantifying the effects of *a*s immediate neighbors on its interest profile, while G_2 is useful in accounting for the cascading effect. For example, the influence then nodes in G_1 can have on *a* can be modeled as a function of its neighbors using the subgraph G_2. These methods of subgraph capture our requirements well and hence, other methods were not explored.

5.2 Interest Profile

The interest profile of a user $I(a)$ is represented as a probability distribution of his interest over the set of categories. Let K be the total number of categories, then $I(a)$ is a K dimensional vector, $[p_k]$, where p_k denote the probability that a user is interested in category k relative to other categories. Here, $k = 1, 2, ., K$

5.3 Interest Influence Model Framework

We hypothesize that the interest profile of a user can be derived from the interest profiles of the followers and that of their friends, using appropriate weighting schemes. Particularly, we propose two models for obtaining a user's interest profile:

$$I_w^{\Phi}(a) = \frac{1}{\sum_{b \in \Phi(a)} w(a,b)} \sum_{b \in \Phi(a)} w(a,b) I(b)$$

$$I_w^{\Psi}(a) = \frac{1}{\sum_{b \in \Psi(a)} w(a,b)} \sum_{b \in \Psi(a)} w(a,b) I(b)$$

where,

- $I_w^{\Phi}(a)$ denotes the predicted interest profile for user a based on $\Phi(a)$ (the set of followers of a) using the weighting scheme w,
- $I_w^{\Psi}(a)$ denotes the predicted interest profile for user a based on $\Psi(a)$ (the set of friends of a) using the weighting scheme w and
- $w(a, b)$ is the weighting factor that captures the influence node b exerts on node a.

This represents the general framework, wherein we use different weighting schemes and obtain different predictions for the interest profile of user. The various weighting schemes implemented are discussed in Section 6.2

5.4 Composite Feature Model

The weighting schemes $w(a, b)$ are so defined (explained in 6.2) to capture a particular aspect of the relationship between a and b. Thus, a number of interest profile predictions can be obtained by using different weighting schemes. Since, each of these predictions uses a single aspect of the relationship, we propose a composite feature model that takes into account all aspects by combining the different predictions in an optimal proportion.

To design such a composite feature model, we require a way of effectively combining the results

from the different weighting schemes. All weighting schemes are not expected to (and from our results, do not) reflect the interest profile equally well. Moreover, different weighting schemes work well for different categories. Thus, we assign a category-wise quality score to each weighting scheme. This quality score is based on how well the weighting scheme identifies the true interest level of the user for that category.

To obtain these quality scores, we formulate a category-wise regression problem. For the regression problem, the independent variables (features / x values) are the probability predictions obtained under different weighting schemes for that category and the dependent variable (predicted variable / y value) is the true interest probability for the user for that category. Solving this regression (curve fitting) problem, gives the coefficients for different features, which represent the quality score for each weighting scheme.

Using these quality scores, we obtain the interest prediction from the composite model as follows-

$$I^c(a) = [i_k^c(a)] \tag{1}$$

$i_k^c(a)$ is the composite predicted interest level for category k, given by

$$i_k^c(a) = \frac{1}{\sum_{w \in W} Q_{kw}} \sum_{w \in W} Q_{kw} i_{kw}$$

where

- W - set of all weighting schemes
- Q_{kw} is the quality score for weighting scheme w for category k
- i_{kw} is the interest level for category k predicted using the weighting scheme w

$I^c(a)$ is normalized to obtain a probability distribution.

5.5 Machine Learning Model

In addition to the graph based interest influence approach described above, we use a machine learning based predictive model to predict the interest of a user using graph features. We use a multi-label, multi-class supervised learning model, where each user is represented by a vector of features obtained from his graph and the variable to predict is the list of all interest categories for the user.

6 Experimental Details

6.1 Obtaining interest profiles

The primary entity of the study is the interest profile of a user. Since the focus of our project is to conduct a graph based interest modeling, we adopt a simple approach to obtain the interest profile and do not explore other options. We use a set of 5 categories - Sports, Entertainment, Food, Health and Technology and map a user's tweets to a probability distribution over these topics. We use a word list of topics and increment the count for each topic if a word of that topic appears in the user's tweet. The distribution thus obtained is normalized to get the probability distribution over categories for that user, which represents the interest profile $I(a)$ for user a.

For example, a tweet - "I love the new iPhone " contains the word "iPhone " which is a part of the words list for the topic "Technology " and hence will output a distribution [0,1,0,0,0]. We obtain these topic-wise word lists by merging the lists from the following sources[5] [6].

6.2 Feature Engineering

As described in 5.2, the predicted interest profile of a user is obtained using a weighted average of the interest profiles of the user's followers/friends. Also, as previously described, we implement a variety of weighting schemes to incorporate different aspects (features) of the relationship between the user and his network.

The features used are listed below. For each feature, we reason the use of the feature and give the weighting scheme w used.

6.2.1 Baseline Weighting

All weights $w(a, b)$ are defined to be 1. Thus, the baseline method does a simple average of all the interest profiles.

6.2.2 Retweet Weighting

A user highly interested in (say) $Technology$ would not only follow a user (say $@TechCrunch$) but is also more likely to interact more with that user. Interactions in Twitter can be captured in terms of ReTweets and Mentions. Thus, we use a Retweet weighting scheme where

$w(a, b)$ = number of times the user a has retweeted a tweet of user b.

[5] http:www.ourcommunity.com.au/tech/tech_article.jsp?articleId=74
[6] http://www.enchantedlearning.com/wordlist/

Since this weighting scheme can potentially have zeros for a lot of users, we use a smoothened measure,

$w(a, b)$ = 1 + number of times the user a has retweeted a tweet of user b

6.2.3 Mentions Weighting

Similar to the ReTweet weighting scheme, we use the mentions to obtain Mentions Weighting, where

$w(a, b)$ = number of times the user a has mentioned b in his tweet.

The smoothened measure is

$w(a, b)$ = 1 + number of times the user a has mentioned b in his tweet.

6.2.4 Mutual Followers Weighting

Another edge based feature is the number of nodes the two nodes are mutually connected to. Twitter graph, being a directed graph, we use two such measures based on in-edges and out-edges.

$w(a, b) = |\Phi(a) \cap \Phi(b)|$

The smoothened measure is $w(a, b) = 1 + |\Phi(a) \cap \Phi(b)|$

6.2.5 Mutual Friends Weighting

This is similar to mutual followers weighting-

$w(a, b) = |\Psi(a) \cap \Psi(b)|$

The smoothened measure is $w(a, b) = 1 + |\Psi(a) \cap \Psi(b)|$

All features described above are edge-based measures. Next ,we list all the node based features we develop. The intuition behind these features is that different nodes have different influencing capabilities and thus, the properties of the relevant nodes for a user influence the interest profile of the user. We implement three such features as given below.

6.2.6 Follower bias

This measures the importance or popularity of the person

$w(a, b) = |\Phi(b)|$

6.2.7 Friend bias

This measures the relevance of the person.

$w(a, b) = |\Psi(b)|$

6.2.8 Tweets bias

This measures the level of activity of the person

$w(a, b)$ = Number of tweets of user b

We use all the weighting schemes described above to obtain both $I_w^{\Phi}(a)$ and $I_w^{\Psi}(a)$.

6.3 Deriving topics of interest from interest profile

Once we obtain the interest profile using these methods, we derived the topics of interest using two methods:

- *Threshold*: Select all topics whose probability (component in the interest profile) is more than a threshold value. We used a threshold of $\frac{1}{K}$, where K is the number of categories. Topics = $\{k|I_k(a) > 1/K\}$
- *ChooseTopM*: Select M topics with the highest probability values. In our experiments, we used a value of $M = 2$, since we deal with only 5 categories.

6.4 Error metrics

To quantify the performance of our models, we used three metrics - Precision, Recall and F1 scores for the true labels and the predicted labels obtained.

6.5 Composite Feature Model

As described in 5.4, we use a regression model to obtain the category-wise quality scores for each weighting scheme. In particular, we used ridge regression where the loss function is the linear least squares function and regularization is given by L2-norm. We identified the regularization coefficient using leave-one-out cross validation. The regression model was then used to arrive at the quality scores. These quality scores where then used in conjunction with the interest profiles obtained from different weighting schemes to arrive at the predicted interest profile using the composite model.

6.6 Machine Learning Model

For the machine learning based prediction task explained in 5.5, we used a Random Forest based classifier and validate the performance using leave-one-out cross validation. The error metric used is the F1 score. Additionally, we used a feature selection algorithm based on the feature importances calculated by the Random Forest classifier to identify the optimal subset of features. The optimal set of features were then used for the prediction task.

7 Results and Analysis

Figures 1, 2 and 3 show the results (Precision, Recall and F1) obtained from the interest influence model for different feature weighting schemes. All results presented here are obtained by averaging the results for 100 users. We observe that the general trend is that the values obtained using the ChooseTopM method are higher than that obtained using the Threshold method. This is attributed to the fact that the ChooseTopM method is agnostic to the probability values and chooses just the topics corresponding to the top M values. Since, we use only a small number of categories in our analysis, it is quite easy for the model to perform better under ChooseTopM method. Contrast this with the Threshold method, which chooses only values that are higher than a threshold. Hence, this method is better indicative of the actual performance of our approaches, given that we use a simple approach based on tweets to identify the interest profiles. Going forward, we use only the Threshold method for analysis.

In the Threshold method, we observe that the follower network gives slightly higher performance in terms of all of precision, recall and F1 than the friend network. While all weighting schemes give comparable performance, only a few weighting schemes give better results. This is partially due to the fact that we use a limited data and a basic approach to topic identification (in terms of text matching) and primarily due to the fact that each feature captures only one aspect of the interest influence.

This leads us on to the composite model, where we combine the features as explained previously. The results obtained using the composite model using both the friend network and follower network are shown in Table 1. The composite model gives a superior performance compared to the base model in terms of F1 score for both the friend and the follower network.

Table 2 shows the results obtained using the machine learning approach. The F1 score obtained here is significantly higher than the interest influence model and the composite model. This indicates that a supervised model that learns from the graph based features leverages the training effectively and gives a higher performance. Thus, using the graph features in conjunction with a machine learning approach is lucrative. Moreover, it is important to note that the dataset used consisted of 100 users and hence, using a larger dataset shall certainly be expected to increase the performance.

8 Discussion

The results obtained using the graph based approaches were encouragingly positive. We dug deep into the results to understand how the performance varies for different users. In particular, we wanted to identify if some graph property of the user was indicative of the performance. Our analysis suggests that clustering coefficient of a user is indeed representative of the efficiency of the graph based model to predict his/her interest profile. Users with high clustering coefficient obtained high F1 scores, while those with lower clustering coefficient values, got low F1 scores. Some representative values are shown in Table 3

9 Conclusion and Future Work

In this project, we have explored various graph features for the task of predicting the interest profiles of a user based on his/her friends and followers. The techniques devised are promising and can definitely be extended to improve the performance. We focused only on the network based features in this study and hence used simple procedures for other aspects. For example, we have used the tweets of the followers/friends of the user to extract the interest profiles of the followers/friends. We could extend this to include the biography of the followers/friends obtained from their Twitter user profile. We have used a naive method to extract the probability distribution of the tweets over the different categories for a user where we directly compare the words in the tweets of the user to the wordlists for each category. This can be improved to check for words with edit-distances of 1 to take into account common spelling errors. This becomes important particularly in Twitter where the language used in informal and error-prone. Moreover, we use a set of 5 pre-defined categories. This can be increased to include more categories.

References

Kwan Hui Lim and Amitava Datta. 2013. Interest classification of twitter users using wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration*, WikiSym '13, pages 22:1–22:2, New York, NY, USA. ACM.

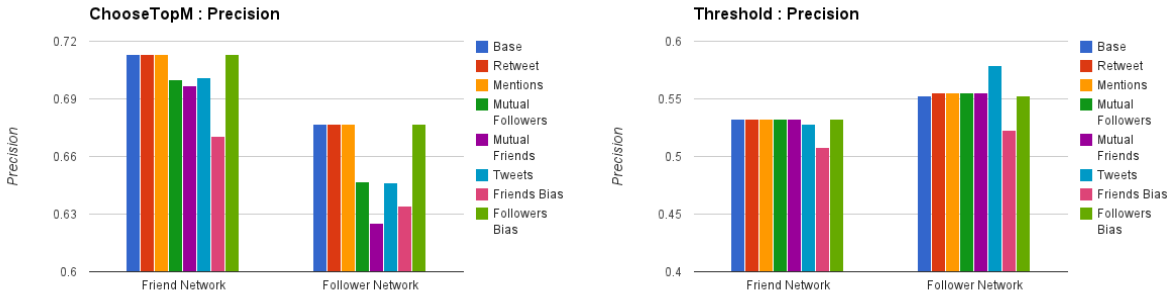

Figure 1: Interest Influence Model : Precision

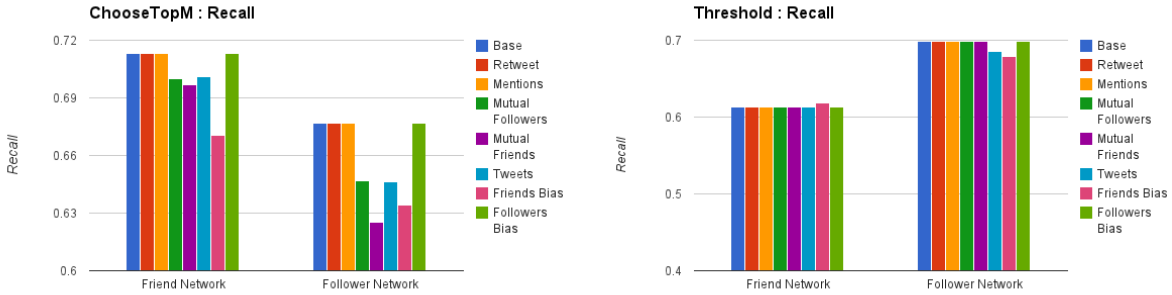

Figure 2: Interest Influence Model : Recall

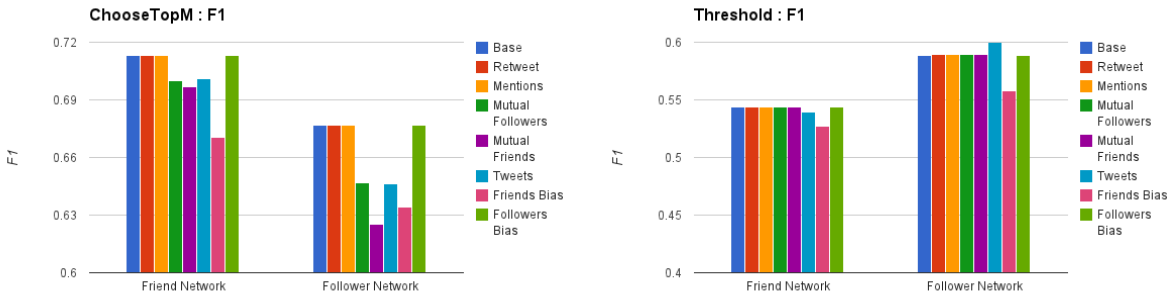

Figure 3: Interest Influence Model : F1

	Precision	**Recall**	**F1**
Baseline - Friend network	0.5325	0.6128	0.5439
Composite - Friend network	0.5966	0.7174	0.6174
Baseline - Follower network	0.5528	0.6992	0.5882
Composite - Follower network	0.6498	0.6908	0.6300

Table 1: Composite Model : Results

Precision	**Recall**	**F1**
0.8079	0.7232	0.7277

Table 2: Classifier Model: Results

Clustering coefficient	**F1**
0.024	0
0.08	0.5
0.41	0.5
1.56	0.6667

Table 3: Clustering coefficient values and F1 scores for sample users

Matthew Michelson and Sofus A. Macskassy. 2010. Discovering users' topics of interest on twitter: A first look. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, AND '10, pages 73–80, New York, NY, USA. ACM.