Detecting User Interests on Twitter via Seed Set Expansion

Amit Goyal Praveen Bommannavar Stuart Anderson Alek Kolcz Kurt Smith

Twitter Inc.
San Francisco, CA, USA
{agoyal,praveen,soa,ark,kurt}@twitter.com

ABSTRACT

A central machine learning problem in user interest modeling is to learn what topics users in a social network (such as Twitter) are interested in. Indeed, this is a critical part of understanding users' behavior. If performed with reasonable precision, it can be exploited in a wide range of applications. Much of the previous work on this problem involves computationally expensive text processing – making the approach language dependent and, thereby, not scalable to international markets. In this work, we propose a novel graph-based approach to interest modeling that is language independent. In particular, we start with a list of people who are influencers and the topics they are KNOWN-FOR (also called *seed set*). First, we expand these KNOWN-FOR topics to other influencers. Later, we learn INTERESTED-IN topics for all users by propagating the known for topics through the social graph. Numerical results show that on the Twitter social network consisting of over 250M users, we are able to grow a seed set of 55K labeled accounts into 88% interest coverage. Additionally, survey results verify that the precision of the detected topics at this coverage is as high as 80%.

Categories and Subject Descriptors H.2.8 [Database Management]: Database Applications - *Data Mining*
Keywords: Social Networks, User Interests, Twitter

1. INTRODUCTION

The objective of the problem of user interest modeling is to learn the topics that users in a social network are interested in. These interest models, if they have reasonable accuracy, can be useful not only for applications to recommendations and search, but in developing fundamental understanding of users' behavior: What kind of users are interested in which topics? How many users are interested in each topic? Which topics are popular in a specific country? What are the growth trends among users interested in various topics? Which topics are growing/shrinking, in terms of active user counts? How do various events impact growth trends in various topic populations? These insights are crucial in making strategic decisions about product and marketing. In this work, we offer a novel user interest modeling technique with the goal of developing insights such as the ones above. To this end, we focus primarily on maximizing the prediction coverage while retaining reasonable precision.

Several attempts have been made on modeling users' interests in social networks, some in the context of Twitter specifically [3, 2, 8, 5, 7, 9]. A common approach in much of the previous work leverages methods for document topic modeling that involve expensive text processing, e.g., leveraging the Tweets/hashtags/mentions they post, profile bio information, search queries, Tweets from the accounts they follow etc. [8, 7]. In the context of Twitter, this is especially challenging as the Tweets are limited to just 140 characters [3]. Further, rather than unlabeled categories of users as one might obtain via LDA or other similar methods [8], we are more interested in assigning interests to users from a known *taxonomy* of several hundred topics. Still another difficulty in the straightforward application of text based methods is the strong dependence on language choice and vocabulary.

In this work, we develop a language independent approach. Our primary focus is to learn the topics that users are interested in. We additionally learn two types of topic assignments: KNOWN-FOR and INTERESTED-IN. While INTERESTED-IN topics are learnt for all users, KNOWN-FOR topics are applicable only to influencers (user having a lot of followers). As an example, Justin Bieber has over 50M followers and is KNOWN-FOR Pop Music. Similarly, Barack Obama has over 40M followers and is KNOWN-FOR Government & Politics. In this work, for the sake of simiplicity, we define the notion of influencer as someone who has 10,000 or more followers. The idea is to distinguish experts/influencers from common users. A follow link from a user to an expert in some topic indicates that the user is interested in the topic. Moreover, while we allow an influencer to have maximum of one KNOWN-FOR label, we learn multiple INTERESTED-IN topics for common users.

While several methods are used at Twitter learn user topic labels, in this work, we highlight an approach that uses a small collection of manually labeled accounts (which we call a *seed set*) with the topics (from a fixed taxonomy) that they are KNOWN-FOR. This database contains around 55K influencers. While this labeled data has very high precision (as it is manually labeled by expert curators), it does not have high coverage: most of the verified accounts are in the USA, UK and Japan. Therefore, our approach is divided into two phases: (i) First, we expand the list of KNOWN-

FOR labels to other influencers using Twitter Lists [1]. (ii) Next, we propagate the topics from influencers to common users via the social graph.

A related problem is of identifying influencers and topical experts [4, 10, 1] in social networks. The idea is to identify and target a small set of influencers for (viral) marketing campaigns. Pal et al. [6] focus on identifying topical authorities on Twitter. While we also in a way, identify influencers and learn KNOWN-FOR labels for them, our primary goal is to detect the user interests and the KNOWN-FOR labels are assigned only as a mean to achieve that goal.

In summary, our contributions are as follows.

- We propose a scalable language independent approach to the problem of user interest modeling. The goal is to maximize coverage, while having reasonable precision (80% or higher). We distinguish and learn two types of topics – KNOWN-FOR for influencers and INTERESTED-IN for all users.
- Through evaluation on Twitter social graph, we show that coverage is quite impressive: almost 90%. To evaluate precision, we conduct user surveys and show that the precision is as high as 80%.

The next section outlines our model and approach. In Section 3, we present an evaluation of our model. Finally, Section 4 concludes the paper.

2. DETECTING USER INTERESTS VIA SEED SET EXPANSION

As mentioned earlier, our algorithm is divided into two phases. In the first phase, we expand the list of KNOWN-FOR labels from the seed set, via Twitter Lists. Then, in the second phase, we propagate the topics from influencers to common users via Twitter social graph. Details are as follows.

2.1 Learning Known-For labels

The KNOWN-FOR labels are learned only for influencers. For the sake of simplicity, we apply a hard constraint – an influencer must have 10K or more followers. Moreover, an influencer can be labeled with only one KNOWN-FOR topic. We begin with the labeled seed set, consisting of 55K influencers and expand it to other influencers via Twitter Lists. A Twitter list is a curated group of Twitter users. A user can create her own list, or follow the lists created by other users. The utility of a list is to have a filtered timeline for individual interest topics. As an example, a user would put Lady Gaga, Justin Bieber and Britney Spears in a list to have a filtered timeline for Pop Music. Similarly, another user may put Barack Obama, Bill Clinton and George Bush in a list to have a timeline for Government & Politics. Thus, these lists can be considered clusters of influencers who are known for similar topics.

Users of Twitter have created millions of such lists. We filter them heavily to obtain a clean set. For instance, we filter out the lists created by users who are estimated to be untrustworthy (spammers, etc.). Similarly, we filter out the lists that contain influencers (in labeled data) known for a very diverse set of topics. We skip these details for brevity.

[1] https://support.twitter.com/articles/76460-using-twitter-lists

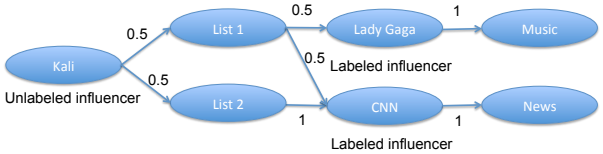

Figure 1: Example: Learning Known-For topics. Arrows represent targeted random walks.

After filtering the lists, we learn the KNOWN-FOR labels as follows. We construct a graph consisting of seed set (labeled influencers), other influencers (users having 10K or more followers) and lists. The edges are drawn among influencers and lists based on memberships. Then, a targeted random walk is started from unlabeled influencers to lists to seed set to topics. Finally, we take the topic with the highest probability (based on random walk) as her KNOWN-FOR topic.

Consider the example shown in Fig. 1. Kali is an unlabeled influencer and a member of two lists. On the other hand, Lady Gaga, who is a labeled influencer, and known for music, is a member of List 1. Similarly, CNN is another labeled influencer account, known for News, is a member of List 2. We start random walk from Kali and compute the probability with which it reaches any of the two topics. It is easy to see that in the example, the probability that a random walk from Kali reaches topic Music with probability $0.5 \cdot 0.5 \cdot 1 = 0.25$, and similarly, reaches topic News with probability $0.5 \cdot 0.5 \cdot 1 + 0.5 \cdot 1 \cdot 1 = 0.75$. Thus, we assign Music as the KNOWN-FOR label to Kali.

This step increases the number of KNOWN-FOR *labels from 55K to 336K.* That is, our algorithm produces an output of 6 times as many experts as we began with. It is worth noting that for some countries this allows us to find experts for a few topics which previously had no coverage.

2.2 Learning Interested-In labels

Once the KNOWN-FOR topics are learnt, we propagate them to all users through Twitter social graph. The idea is similar: we perform random walks from users to influencers (that are labeled, after the first step). A key difference in both phases is that while we restrict an influencer to be known for only one topic, we allow users to be interested in multiple topics. For simplicity, we also reset the KNOWN-FOR weights to 1.

In our analysis, we found that allowing multiple hops in random walks decreased the precision. Hence, we restrict the random walks to one hop. This approach is especially well suited for the Twitter graph, which is a relatively flat social network (anyone can follow anyone). Hence, *explicit* follow links indicate the cleanest form of users' interest.

Consider the example shown in Fig. 2. Here, a common user Emily follows 3 influencers, for whom we now have the KNOWN-FOR labels. We start a random walk from Emily, and compute the probability with which it reaches an interest topic. That probability is then inferred as weight with which the user is interested in the given topic. As an example, Emily reaches Music with probability 0.33 and News with probability 0.67. Thus, we infer that she is interested in Music with weight 0.33 and in News with weight 0.67. Note that we reset all KNOWN-FOR weights to 1 after first step.

2.3 Further Improvements

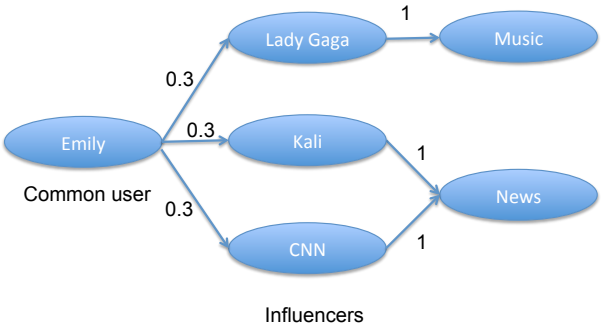

Figure 2: Example: Learning Interested-In topics. Arrows represent targeted random walks.

There are two issues we observe upon inspecting INTERESTED-IN topics detected by the above method. First, this method is prone to *overfitting* in the case when a user follows only one influencer. Another issue with this approach is the limited coverage. In particular, this method provides us the coverage of 78%. That is, we can detect one topic for 78% of monthly active users (there are over 250M monthly active users). We address these issues by performing a 2-hop random walk, instead of 1-hop. Note that the 2-hop random walk is performed only from the users who are either a) not covered in the 1-hop random walk, or b) may lie in the overfitting case. This improvement increases both the precision (by smoothing the overfitting issue mentioned previously) as well as the coverage. In particular, coverage increases from 78% to 88%. More details are provided in Section 3.

3. EVALUATION

In this section, we evaluate our seed set expansion methodology on Twitter social graph on two grounds: a) coverage of the resulting INTERESTED-IN topics assigned to users, b) precision of the assigned interest labels as measured via a user interest survey. The graph contains over 250M users (we do not provide the exact numbers due to proprietary).

3.1 Coverage

Overall coverage achieved by our algorithm is 88% – that is, 88% of users are labeled with at least one topic. This is impressive, as our algorithm is applicable to all users, irrespective of the language they speak. On the other hand, most previous works [3, 8, 5, 7, 9] are constrained by the languages.

The distribution of maximum interest weights (or scores) is shown in Fig. 3. On X-axis, we have the maximum interest weight (or score) that a user gets for any INTERESTED-IN topic and on Y-axis, we have fraction of such users. For example, the maximum weight for 20% of users is in [0.2, 0.3). As mentioned above, 88% of users are labeled with at least one topic (maximum weight for 12% users is 0).

3.2 Precision

We evaluate the precision of the detected user interests by conducting a survey (interface shown in Fig. 4). Users were presented a list of topics with the question: `I would like to see Tweets about this topic`. We used a 7-point scale for registering the responses. In addition, users could also choose `I don't understand this topic`.

Users who did not complete any part of their survey

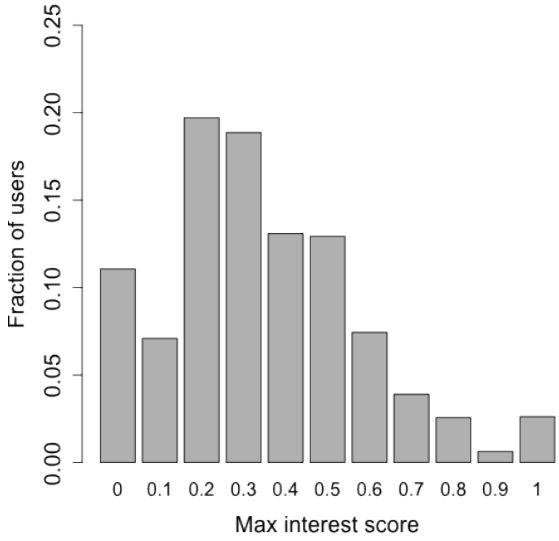

Figure 3: Fraction of users vs. maximum interest score.

Figure 4: Surveying users to evaluate detected Interested-In topic precision.

or marked `I don't understand this topic` were removed from evaluation. That left us with responses from 2804 users. In the remaining seven point scale, those entries marked with `Strongly agree`, `Agree` or `Somewhat agree` were counted as true positives, while the rest were counted as false positives.

Fig. 5 shows the achieved precision of a detected topic given the weight (or score). *As can be seen, the precision of the detected topics is consistently above 80%.* The figure verifies that the precision is healthy irrespective of the interest weights.

Due to the fact that many users do not receive topics with scores in the range of 0.9 to 1.0, our survey sample resulted in some bins having less than 50 samples each. Therefore, we discard those bins and focus instead on the most common bins. Despite this missing data, the overall precision estimate for this data is relatively stable, since the missing buckets account for a small fraction of the population of users.

One can observe that the score is not a strong predictor of precision - this is because this score also depends on the total number of users with KNOWN-FOR labels that are being

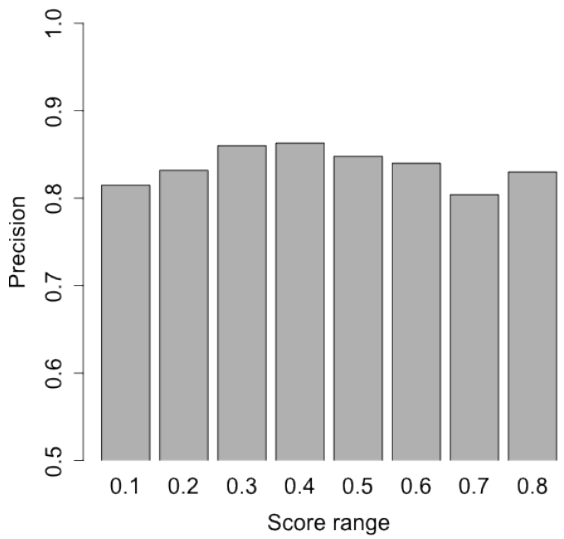

Figure 5: Precision vs. topic score.

followed.

In addition to measuring the precision of our predicted user topics, we also measured the precision of two baseline methods. First, choosing the three most common interests (`celebrities`, `sports`, `music and radio`) and randomly assigning them to users results in a precision of 57%. Randomly assigning any of the topics in our taxonomy to uses yields an far worse precision of 29%.

We note that there exists some amount of unavoidable bias in our survey. Despite making it available for several days, it is possible that only a certain type of user would even respond to it, making it difficult to measure precision on a truly uniform sample of users. For the type of descriptive analysis mentioned in this paper, we posit that this measurement bias is tolerable.

3.3 IPL Cricket Season: A Case Study

Next, as a case study, we consider the Indian Premier League (IPL) Cricket season (the one in 2013) by looking at its impact on users' interests. This Cricket tournament was held in 2013 from April 3 to May 26. We tracked the users' interests in Cricket using our model from Sept 2012 to Aug 2013. The results are shown in Fig. 6. In this plot, we define the interest share as sum of the interest weights of users in Cricket in India. Then, we normalize these interest shares on the value of Sept 2012. As can be seen from the plot, we observed a substantial increase in the interest share during the window of Cricket season. In particular, the interest share in Cricket jumped to 92% in the month of April (relative to what it was in Sept 2012).

4. CONCLUSION

In this paper we have introduced a technique for leveraging a relatively small set of topical expert labels into user interest predictions for a large fraction of users on Twitter. In the first stage, the KNOWN-FOR labels are expanded to a larger set of authoritative accounts by examining co-occurrences in Twitter Lists. The next stage propagates these labels into scored INTERESTED-IN assignments for non-authoritative and authoritative users alike. Almost 90% of users on Twitter can be assigned a top interest in this manner, and the precision of the labels has been shown in evaluation surveys to be at least 0.80.

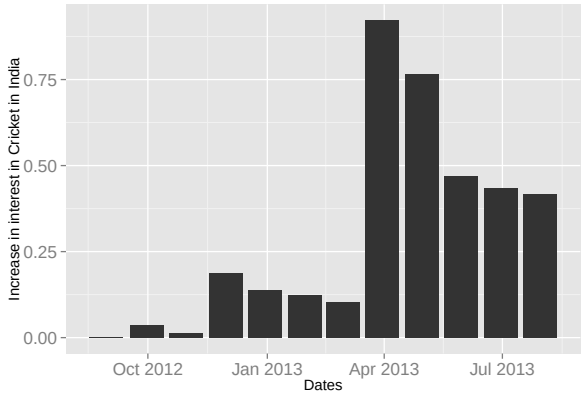

Figure 6: Increase in interest share in Cricket in India during IPL Cricket Season.

5. REFERENCES

[1] G. S. Bevilacqua, S. Clare, A. Goyal, and L. V. S. Lakshmanan. Validating network value of influencers by means of explanations. In *IEEE International Conference on Data Mining*, ICDM '13, 2013.

[2] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10. ACM, 2010.

[3] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *WebKDD and SNA-KDD Workshop on Web Mining and Social Network Analysis*. ACM, 2007.

[4] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proc. of the Ninth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'03)*.

[5] D. Kim, Y. Jo, and I. chul Moon. Analysis of twitter lists as a potential source for discovering latent characteristics of users. In *ACM CHI Workshop on Microblogging*, 2010.

[6] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *ACM International Conference on Web Search and Data Mining*, WSDM '11, 2011.

[7] M. Pennacchiotti and A.-M. Popescu. A machine learning approach to twitter user classification. In *ICWSM*, 2011.

[8] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.

[9] N. K. Sharma, S. Ghosh, F. Benevenuto, N. Ganguly, and K. Gummadi. Inferring who-is-who in the twitter social network. In *ACM Workshop on Workshop on Online Social Networks*. ACM, 2012.

[10] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *ACM International Conference on Web Search and Data Mining*, WSDM '10. ACM, 2010.