Topic-Association Mining for User Interest Detection

Anil Kumar Trikha, Fattane Zarrinkalam, and Ebrahim Bagheri

Laboratory for Systems, Software and Semantics (LS^3), Ryerson University
{atrikha, fzarrinkalam, bagheri}@ryerson.ca

Abstract. The accurate identification of user interests on Twitter can lead to more efficient procurement of targeted content for the users. While the analysis of user content has engaged with on Twitter is a rich source for detecting the user's interests, prior research have shown that it may not be sufficient. There have been work that attempt to identify a user's *implicit interests*, i.e., those topics that could interest the user but the user has not engaged with them in the past. Prior work has shown that *topic semantic relatedness* is an important feature for determining users' implicit interests. In this paper, we explore the possibility of identifying users' implicit interests solely based on topic association through frequent pattern mining without regard for the semantics of the topics. We show in our experiments that topic association is a strong feature for determining users' implicit interests.

1 Introduction

User interest detection techniques that automatically identify users' interests towards active topics on Twitter have become an emerging research area in the recent years, primarily due to its potential to improve the quality of higher-level applications such as news recommendation [1] and retweet prediction [2], among others. Most of the existing work in the field of user interest detection are focused on extracting explicit interests via analyzing textual contents shared by the users [4, 12]. Based on the fact that the majority of users in social networks are not very active (free-riders), their available content is sparse and does not reveal sufficient clues about their interests. To address this challenge, there have been work dedicated to inferring implicit interests of users [7, 9]. Implicit interests are those potential interests that might be relevant and interesting for the user but the user has not engaged in them explicitly in the past [11].

Several authors have indicated that interaction patterns between users and topics are among the important clues for determining implicit interests [8, 10]. To systematically investigate the suitability of users' interaction patterns and topic relatedness on the quality of implicit interest detection, a graph-based link prediction scheme is proposed in [11], which combines these two factors into a unified representation model. Based on the experiments, the authors found that topic relatedness is a contributing factor that can accurately uncover implicit interests of users. In other words, users on Twitter are more likely to be interested in topics that are conceptually similar to the topics that they have explicitly engaged with in the past. On the basis of this finding, in this paper, we are interested in topic association as a means to infer implicit interests of users by turning the implicit interest detection problem into a frequent pattern mining problem. Frequent pattern mining (FPM) is a widely adopted data mining technique that has

mostly contributed to the discovery of co-occurrences and associations between items of a dataset. FPM methods have already been used in the field of social network analysis for finding hidden patterns in social data [5, 6]. In line with these works, we apply FPM in the context of implicit interest detection to extract the association between topics on Twitter and subsequently infer users' implicit interests. We then build the interest profile of a user considering both her explicit and implicit interests.

2 Proposed Approach

We study the problem of inferring user interest profiles towards active topics on Twitter, within a given time interval, which can formally be defined as follows:

Definition 1 ***(Interest Profile)*** *Given a set of K topics $\mathbb{Z}$, an interest profile of a user $u \in \mathbb{U}$ in time interval T, called $P^T(u)$, is represented by a vector of weights over K topics, i.e., $(f_u(z_1), ..., f_u(z_K))$, where $f_u(z_k)$ denotes the degree of u's interest in topic $z_k \in \mathbb{Z}$ at time interval T. A user topic profile is normalized using $L1 - norm$.*

Our proposed approach performs the following three steps to infer the interest profile of users: *1)* Inferring users' explicit interests by extracting information from the content that the users have shared on Twitter; *2)* Generating frequent patterns based on the collective set of users' explicit interests in order to understand the relation between topics in a given time interval T; and, *3)* Augmentation that incorporates additional implicit interests into a user's interest profile based on the frequent patterns learnt in Step 2. These three steps are described in the following.

2.1 Inferring User Explicit Interests

The interests that are observable in a user's tweets are known as *explicit* interests. User explicit interest detection methods from Twitter have been studied in the literature and therefore are not the focus of our work and we are able to work with any topic and interest detection method to extract topics $\mathbb{Z}$ and the explicit interest profile of each user u toward these topics in time interval T, denoted as $P_E^T(u) = (f_u^E(z_1), ..., f_u^E(z_K))$. Considering $\mathbb{M}$, the set of available microposts, it is possible to extract topics $\mathbb{Z}$ using Latent Dirichlet Allocation (LDA), the *de facto* standard in topic modeling. As suggested in [11, 12], to obtain better topics from Twitter without modifying the standard topic detection methods, we annotate the text of each tweet with Wikipedia concepts using an existing semantic annotator. Next, given the published or retweeted microposts of a user u, $\mathbb{M}_u$, we initially divide $\mathbb{M}_u$ into N segments based on a uniform time interval T, $\mathbb{M}_u = \{\mathbb{M}_u^1, \mathbb{M}_u^2, ..., \mathbb{M}_u^N\}$. Then, we aggregate all concepts extracted from each tweet segment of a user into a single document and apply LDA on the collection of such documents to discover K topics $\mathbb{Z}$, and explicit interest profile of each user u in each time interval T, i.e., $P_E^T(u)$.

2.2 Discovering Frequent Topic Patterns

Given the collective set of users' explicit interests, i.e. $\{P_E^T(u) | 1 \leq T \leq N, u \in \mathbb{U}\}$, in this section, we aim at utilizing FPM methods to find closely related topics that frequently

co-occur within the explicit interests of our user set. To do so, we treat topics $\mathbb{Z}$ as items and use the explicit interest profile of each user in a given time interval T, $P_E^T(u) = (f_u^E(z_1), \ldots, f_u^E(z_K))$, to form a transaction τ. Thus each transaction τ consists of the set of topics that a user is explicitly interested in at time T, i.e., $\tau = \{z | f_u^E(z) > 0\}$. Then, we apply an FPM method to mine the transactional database built based on Definition 2 to calculate the frequent topic patterns in time interval T, denoted as $FP_{\mathbb{Z}}^T$.

Definition 2 ***(Transactional Database)*** *The transactional database for time interval T, denoted as TDB^T, includes the collective set of all users' explicit interests in time interval T and L past time intervals, i.e. $TDB^T = \{P_E^t(u) | T - L \leq t \leq T, u \in \mathbb{U}\}$.*

In Definition 2, in order to be able to study the impact of considering historical user interests on the performance of extracted frequent topic patterns, we also add the historical explicit interest profile of all users in L past time intervals to the transactional database. Appropriate algorithms like *Apriori*, *Eclat* and *FP-Growth* have been developed to efficiently discover frequent patterns. In this work, we utilize the FP-Growth algorithm as an efficient method which mines frequent patterns without costly candidate generation. It has been experimentally shown in [3] that FP-Growth algorithm has the best performance among the others and is thus the most scalable.

Now, given the explicit interest detection method described in Section 2.1 is based on dividing the user's data into N discrete time intervals, we perform the above process for each of the intervals T. This will produce $\{FP_{\mathbb{Z}}^T | 1 \leq T \leq N\}$, which is the input of our augmentation method to build interest profile of each user in each time interval T. For example, $s = \{z_{35}, z_{88}\}$ is the most frequent topic-set extracted for December 10, 2010. Topic z_{35} = {`Mixtape`, `Hip_hop_music`, `Rapping`, `Kanye_West`, `Jay-Z`, `Remix`} refers to the hip-hop music collaboratively produced by American rappers *Jay-Z* and *Kanye West* and topic z_{88} = {`Lady_Gaga`, `Song`, `Album`, `Concert`, `Canadian_Hot_100`} refers to the concert of *Lady Gaga* in Canada. It is clear that these two topics are related to music and the users who are explicitly interested in z_{35} could potentially also be interested in z_{88}.

2.3 Interest Profile Augmentation

In this section, to build the interest profile of a user u in time interval T, $P^T(u)$, as defined in Definition 1, we augment the explicit interest profile of the user $P_E^T(u)$, using $FP_{\mathbb{Z}}^T$, the frequent topic patterns in time interval T, based on Algorithm 1. As shown in the Algorithm, given $P_E^T(u) = (f_u^E(z_1), ..., f_u^E(z_K))$, we take each topic z which is of explicit interest to user u, i.e, $f_u^E(z) > 0$, and search $FP_{\mathbb{Z}}^T$ to find any topic-set s which includes topic z (Lines 3 to 5). If such a topic-set s exists, we take the other topics in s and add them to the interest profile of user u, $P^T(u)$ (Line 7). At the end of this process, the explicit interest profile of each user in each time interval is augmented with additional interests from the frequent topic patterns.

3 Experiments

We use the publicly available Twitter dataset [1] that includes 3M tweets posted by approximately 135K users, starting from Nov. 1st and lasting for two months until Dec. 31st

2010. As mentioned in Section 2.1, we annotated the text of each tweet with Wikipedia concepts using the TagMe RESTful API, which resulted in 350,731 unique concepts. Then, we applied the Gensim implementation of LDA to extract topics and explicit interests of users over these topics in each time interval T. The number of topics is set to 100 and the length of time interval T is set to 1 day.

Evaluation Methodology and Metric. Adopted from [9], we deploy a retweet prediction application for evaluation. Since the main goal is not to propose a retweet prediction system, the authors have adopted a simple algorithm which is only based on user interest profiles. To do so, given the tweets of two consecutive time intervals, i.e., T_1 and T_2, for a user u, her interest profile $P^{T_1}(u)$ is built based on the tweets that she has published or retweeted in time interval T_1. Further, the tweets that she has retweeted in time interval T_2 are considered to be the ground truth for that user in order to evaluate the results of the retweet prediction application. For user u, to predict a retweet, the tweets of her followed users from whom she has retweeted at least one tweet in time interval T_2 are considered as candidates, and the topic similarity between a candidate tweet and the user interest profile of user u is computed as described in [9].

Then, we rank the tweets based on the similarity scores in descending order. By comparing the ranked list of candidate tweets with the ones that are in the ground truth, we evaluate the quality of retweet prediction, and therefore determine how successfully the interests of a user have been identified. We adopt Mean Average Precision (MAP) as our evaluation metric.

Comparison Methods. we consider the following user interest detection methods for comparison: (1) **EUI**: In this method, the **E**xplicit **U**ser **I**nterest detection method described in Section 2.1, is used to build user interest profiles. (2) **Zarrin's Model**: This method builds user interest profiles based on combining Explicit and Implicit Interest profiles. In this method, we build $P^T(u)$, by augmenting $P_E^T(u)$, explicit interests of user u at T, with the implicit interest of user u to each topic z that she is not explicitly interested in, i.e., the value of $f_u^E(z_k)$ is equal to 0. To infer the implicit interests, we follow the proposed link prediction method described in [11]. Based on the results in [11], we selected the best configuration of this paper i.e., S that considers the semantic relatedness between topics and Adamic/Adar as link prediction method, for comparison here. (3) **Wang's Model**: This method which is proposed by Wang et al. [9] learns interest profile of user u, i.e., $P^T(u)$, based on a link structure regularization framework that consider both user explicit interest and the relationship between users to detect implicit interests.

3.1 Effect of Parameters

By setting the value of the minimum support threshold *minsup* in the frequent pattern mining process, it is possible to generate variable number of patterns as needed. Further, as described in Section 2.2, L denotes the number of historical time intervals included in the transactional database to extract frequent topic patterns in each time interval. Here, we investigate the impact of these parameters on the quality of our proposed method by changing the value of *minsup* from 0.4% to 4% and the size of L from 0 to 5. The results are reported in Figure 1. Based on results, the quality of prediction results in terms of MAP has significantly decreased by increasing the value of minsup value from 0.4% to 4%. When minsup is set low, the number of frequent topic-sets increases dramatically.

Algorithm 1 Augmentation process

Input: $P_E^T(u), FP_{\mathbb{Z}}^T; 1 \leq T \leq N$
Output: $P^T(u); 1 \leq T \leq N$

for $P_E^T(u) = (f_u^E(z_1), ..., f_u^E(z_K)) : 1 \leq T \leq N$ **do**
  $P^T(u) \leftarrow P_E^T(u)$
  for $z \in \{z | f_u^E(z) \geq 0\}$ **do**
    for $s \in FP_{\mathbb{Z}}^T$ **do**
      if $z \in s$ **then**
        for $x \in s$ **do**
          $f_u(x) \leftarrow 1$
return $P^T(u)$

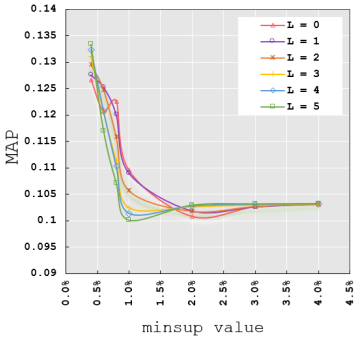

Fig. 1. Effect of the value of minSup and L on the performance of the proposed model.

Thus, it can be concluded that increasing the number of frequent topic patterns leads to user interest profiles that are richer for predicting relevant tweets to a given user. As another observation, it can be seen that considering the historical data of users does not have a significant impact on the increase or decrease of the quality of prediction results. This means that to infer the interest profile of users in each time interval, considering the information provided by users in that time interval is adequate to extract the relatedness between topics. Therefore, in the rest of our experiments, we set the minsup value to 0.4% and L to 0 in our model.

3.2 Comparison With Baseline Methods

We compare the quality of our predicted results with the results of comparison methods in terms of MAP. The results are reported in Table 1. The EUI model is only based on explicit interests of users. Based on the results, it can be observed that it performs worse than the other methods which are all based on both users explicit and implicit interests. This means that incorporating user implicit interests in addition to their explicit interests leads to user profiles that are more accurate for predicting relevant tweets to a given user. In other words, the content generated by users does not reveal sufficient clues to extract all of the users' interests. Therefore, the incorporation of the indirect association between topics or relationships between users can lead to a more accurate representation of users' interests and consequently improve the quality of recommendations.

Based on the results, Both Zarrin's and our model which utilize some form of association between topics to extract implicit interests of users outperforms Wang's model that utilizes the relationship between the users. In line with results reported in [11], this can indicate that finding topic association has a higher influence on identifying users' implicit interests as opposed to considering users' social connections. Based on the Zarrin's model, a user is interested in topics that are conceptually similar to the topics that they have explicitly engaged with. Therefore, it calculates the semantic relatedness between topics based on their constituent concepts and then applies link prediction to infer implicit interests of each user. However, in our proposed model, given the explicit interests of all the users, the implicit interest detection problem is converted into a frequent

Table 1. Performance comparisons. * shows significant difference over baselines at p-value<0.01.

Mehtod	EUI	Zarrin's Model	Wang's Model	Our Model
MAP	0.078	0.096	0.080	0.134*

pattern mining problem to extract relationships between topics. As shown in Table 1, our model builds more accurate user profiles which contribute to improved quality of retweet prediction. This shows that frequent pattern mining methods that do not consider the semantics of topics and only focus on topic co-occurrence can also capture topic association to an accurate degree.

4 Conclusion and Future Work

In this paper, we proposed an approach for identifying user interests over a set of topics on Twitter, considering both their explicit and implicit interests. We model the problem of inferring implicit interests as a frequent pattern mining problem to extract the association between topics and subsequently augmenting explicit interests of users. As future work, based on the fact that users are interested in topics that are conceptually similar, we intend to include semantic similarity between topics in our framework, and infer interest profile of users considering both association and semantic similarity of topics.

References

1. F. Abel, Q. Gao, G. Houben, and K. Tao. Analyzing user modeling on twitter for personalized news recommendations. In *UMAP*, pages 1–12, 2011.
2. W. Feng and J. Wang. Retweet or not?: personalized tweet re-ranking. In *WSDM*, pages 577–586, 2013.
3. K. Garg and D. Kumar. Comparing the performance of frequent pattern mining algorithms. *International Journal of Computer Applications*, 69(25):21–28, May 2013.
4. P. Kapanipathi, P. Jain, C. Venkatramani, and A. P. Sheth. User interests identification on twitter using a hierarchical knowledge base. In *ESWC*, pages 99–113, 2014.
5. S. A. Moosavi, M. Jalali, N. Misaghian, S. Shamshirband, and M. H. Anisi. Community detection in social networks using user frequent pattern mining. *Knowl. Inf. Syst.*, 51(1):159–186, 2017.
6. G. Petkos, S. Papadopoulos, L. M. Aiello, R. Skraba, and Y. Kompatsiaris. A soft frequent pattern mining approach for textual topic detection. In *WIMS*, pages 25:1–25:10, 2014.
7. G. Piao and J. G. Breslin. Inferring user interests for passive users on twitter by leveraging followee biographies. In *ECIR*, pages 122–133, 2017.
8. W. Shen, J. Wang, P. Luo, and M. Wang. Linking named entities in tweets with knowledge base via user interest modeling. In *KDD*, pages 68–76, 2013.
9. J. Wang, W. X. Zhao, Y. He, and X. Li. Infer user interests via link structure regularization. *ACM TIST*, 5(2):23:1–23:22, 2014.
10. Z. Wen and C. Lin. Improving user interest inference from social neighbors. In *CIKM*, pages 1001–1006, 2011.
11. F. Zarrinkalam, H. Fani, E. Bagheri, and M. Kahani. Inferring implicit topical interests on twitter. In *ECIR*, pages 479–491, 2016.
12. F. Zarrinkalam, H. Fani, E. Bagheri, M. Kahani, and W. Du. Semantics-enabled user interest detection from twitter. In *WI-IAT*, pages 469–476, 2015.