Modeling Temporal Dynamics of User Interests in Online Social Networks

Bo Jiang¹, Ying Sha¹

Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{jjiangbo,shaying}@iie.ac.cn

Abstract
Recent years have witnessed an explosive growth of Online Social Networks (OSNs), which serve as a fertile ground for research such as, characterizing individual and group behaviors, identifying information diffusion patterns, and building new recommendation system. This paper explores user interests in social network. While user interests has been extensively studied as the fundamental solution, it neglects the point that a user may change her interests due to social status shift over time. In this paper, we explore two main problems: how user interests change over time and whether user interests have hierarchy. To this end, we first formulate the user interests problem, then adopt semantic enrichment method to determine user interests, and finally employ the topic hierarchy tree model to capture user interests change over time and identify interest hierarchy. Experimental results demonstrate user interests can be divided into primary interest and secondary interest. the primary interest of user hold stability in a long-term period; the secondary interest, however, is more likely to keep up with hot topics or events in the moment. Meanwhile, We also test and compare our model with two existing systems - Who likes what? and TUMS, the result shows that our model can be profiled a more fine-grained user interests.

Keywords: online social network, user interests, semantic enrichment, time-sensitive

1 Introduction

With the advent of the Internet, mobile platforms, online commerce, and social media services, the footprints of human behavior are easily recorded in the digital world, generating data on an extremely large scale. In particular, with the popularity of online social networking platforms and social computing systems, we have witnessed significant growth in generation and usage of social-enriched data in recent years. Social enriched data are becoming increasingly crucial in many aspects of our lives. In the scope of personal activities, the data help us to find interesting topics, parties that friends are attending and news that the world have been happening. One can easily enjoy recommendations and other services tailored for him/her based on the social information. Empowered with personal activities on daily basis and their social context, such
data can provide an insightful view into trends in user interests and behaviors, thus lead to better decision making for governments and businesses.

In order to construct a personalized recommendation for a given user, it is first necessary to have some knowledge about that previous user interests and behaviors. The profile of user interests can be built by considering personal information and contents made by the user. In early work these interests were constructed using the raw terms of prior the user post content, usually in the form of language models, however this often proves to be ineffective, perhaps because such a representation of interests is only considered for a static process. In fact, people's social status will shift with time, accompanied by, their interests also will change over time. So, past posts and recently posts are viewed as equal importance which are not necessarily compatible. In general, the fresh interests seem to better reflect current user demands.

In this work, we attempt to address this shortcoming of static interest by providing a framework for identifying user interests change over time based upon the tweets posted by the user on Twitter. For example, a user always posted tweets about machine learning in the past, while he now is interested in deep learning. We can say that his interest is changed. Once the change is identified as a new interest, a recommendation system can exploit it combining with other user’s interests to personalize user experience and recommend content. We take Twitter as our object due to (i) easy access to data, and (ii) [2, 11] have been concluded that tweets are a good indicator for profiling user interests. For inferring topics of interest of users, we use external knowledge sources as the knowledge base to construct a Topic Hierarchy Tree (THT). The inferred user’s interests are profiled in a tree form. This profile can be utilized for personalized recommendation system with flexibility filter content based on topics of interest of users.

The major contribution of this paper is summarized as follows:

• We propose a method for creating topic hierarchy tree of user interests using external knowledge sources. We found that user’s topics of interest are clearly obtained in this manner.

• We also propose a method for exploiting concepts to represent user interests, scoring confidence of concepts and ranking the importance of user interests.

• We demonstrate the user interests can be divided into primary interests and secondary interests. the primary interests of user hold stability over time; the secondary interests, however, is more likely to keep up with hot topics or events in the moment. We also give a certain interpretation.

The rest of the paper is organized as follows. Section 2 surveys the previous work on analysis of user interests in social networks. In section 3, we show the proposed method of generating user interests using linked data as user interests representation. Section 4 presents the collect method of dataset and gives an evaluation of user interests change over time. The empirical experiment results are reported in Section 5. At last, we conclude our work and give the future research directions in Section 6.

2 Related Works

There has been a number of efforts that study the properties of user interests in OSNs. In light of user interests modeling, we distinguish between three main groups: approaches exploiting
local features, and approaches exploiting the link structure of knowledge sources (e.g., dbpedia or freebase) for semantic linking, and approaches exploiting multiple social networks.

In the first class of approaches, most prior studies attempt to infer topics of interest from the contents of tweets [3, 5, 7, 18, 19, 20]. Weng et al. [18] collect the tweets published by Twitter users into a big document and employ LDA to discover his latent topics of interest. Chen et al. [5] compare two different bag-of-words profiles for each Twitter user and find that profile built on user’s self-tweets work better than on her followings’ tweets. However, Bhattacharya et al. [3] propose a novel methodology to infer topics of interests of users in the Twitter social network. Comparison with topics extracted by content-based techniques reveals interesting insights, [3] finds that the tweets which a user receives are a better indicator of her interests, than the tweets she herself posts. Hong et al. [7] address the problem of predicting user decisions and modeling user interests by analyzing information gathered from Twitter, such as tweets, co-followees, co-followers with factorization matrix.

For the second class of approaches, it is interesting to exploit the link structure of knowledge sources [1, 2, 8, 10, 14, 16]. Abel et al. in their work [1, 2] combine Twitter messages with related news articles and exploit semantics extracted from news articles to infer and contextualize the semantic meaning of each of tweets. Tao et al. [16] use the same method to deduce semantic user interests from tweets, and develop a web service’s TUMS system. In those work, Wikipedia has been widely leveraged for semantics user interests modeling tasks. Kapanipathi et al. [8] has been utilized Wikipedia graph with an adaptation of spreading activation theory to determine the hierarchical interests of users based on their tweets. Michelson and Macskassy [10] propose a model that identifies topics of interest of Twitter users based on their tweets. Their approach mainly relies on extracting and disambiguating the entities mentioned by a tweet. Finally, each entity is assigned the most likely topic retrieved by a sub-tree of Wikipedia categories.

Beside, some previous works have also studied cross-OSN user interests [9, 11, 12]. Ottoni et al. [12] studied the topics of interest for an individual user across Pinterest and Twitter. The work is merged into a supervised learning context where each topic corresponds to a label of the corpus in a one-to-one correspondence (e.g., Labeled Latent Dirichlet Allocation [13], which is a supervised version of LDA) to infer user interests. Orlandi et al. [11] generate semantic user interests used by the weighting scheme, and provide an aggregated score for concepts from multiple social networks (Facebook and Twitter) with a temporal decay.

The above idea of using local social feature and content feature of a user with a recommendation engine to infer a user interests has been around for a long time, and there is a few significant improvement in ways that the problem has been tackled. Nonetheless most of the existing work rely on the assumption that user interests remains static over time and the degree of importance of each interest is equal. This is not consistent with the actual situation.

Our work is inspired by the second kind of work. Meanwhile, because short-term data is often too sparse to allow for robust method performance, we thus focus on long-term tweets data to construct user interests as it provides a richer source of information about the user interests and preferences. We now present an approach to infer temporal-based user interests combing semantic technologies and long-term of user behaviors.

3 Building User Interests Model

The objective of this work is to categorize tweets into a Topic Hierarchy Tree (THT) that constructs a model of user interests by exploiting external knowledge sources. To accomplish this our model as illustrated in Fig. 1, performs the following steps: (1) construct Topic Hierarchy Tree (THT) that is employed multiple knowledge sources (2) User Interests Generator
(UIG) generates topics of interest of user from the tweets of Twitter user. The module first identify entities that are can be directly extracted from user’s tweets and then scores them to reflect the weight of each topic of interest of user. (3) Each of user’s tweets will be added as leaf nodes to the Topic Hierarchy Tree by linking to the most likely topic categories.

![Figure 1: Framework of user interests generator using linked data](image)

3.1 Topic Hierarchy Tree

We utilize external knowledge sources (e.g., dbpedia, freebase, yago, website) as the knowledge-base for construct Topic Hierarchy Tree (THT), as in Figure 2. Topic Hierarchy Tree (THT) extends to over 1000 topics categories and basically includes all aspects of what people talk about. This taxonomy defines contextual topic categories for up to 5 different levels. The first level is a broad level category commonly used as a root for either category or page. Level 2 categories and greater are additional categories nested under Level 1 categories. We exploit Topic Hierarchy Tree to classify user’s tweets into its most likely topic category up to five levels deep. Deeper levels allow to classify tweet into more accurate and more fine-grained topic categories. For instance, an application focused on identifying content discussing personal lending practices can narrow its classification into sub topics (e.g., home financing, family loans, personal loans, student loans) that target decisions with finer resolution.

![Figure 2: An overview of topic hierarchy tree](image)

3.2 User Interests Generator

To better understand the temporal dynamics of user interests and concern how the topics of interest of individual user change over time, we propose to model user’s interests on Twitter
as a set of weighted concepts where a concept may refer to an arbitrary entity and the weight indicates how important the concept is for user’s interests.

Definition 1 (User Interest). The topics of interest of a user $u$ is a set of weighted concepts where a concept $c$ is represented via a named entity.

$$I(u, t) = \{(c, w(c, t, \text{tweet}_u)) | c \in C_E\}$$  \hfill (1)

Where $w(c, \text{time}, \text{tweet}_u)$ is a weight which is computed for the concept $c$ by the given user $u$ based on tweets $\text{tweet}_u$ posted by $u$ and based on the given time $t$. $C_E$ is consisted of a set of entities.

Learning Genuine Interest [6] shows that extracting meaningful concepts only from the tweets is a relatively difficult task, due to their short, noisy, context dependent, and dynamic nature. For instance, given a tweet such as "#Gravity is beautiful http://fb.me/3SkiI5DND" it is difficult to understand to talk about movie or space. However, the semantics of the tweet can be understood by following the link posted in the tweet. Therefore, we first identify concepts from a user’s tweets by Named Entity Recognition, and then extend more enrich semantics.

A great deal of Named Entity Recognition techniques and systems have been developed in tweet [15, 17]. Since our goal is on task of user interests model and not develop entity recognition method, we finally employ an existing solution. To this end, we use TwitIE [4] for our work because the system has relatively superior performance to other services as shown in Table 1.

<table>
<thead>
<tr>
<th>System</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANNIE</td>
<td>47%</td>
<td>83%</td>
<td>60%</td>
</tr>
<tr>
<td>TwitIE</td>
<td>77%</td>
<td>83%</td>
<td>80%</td>
</tr>
<tr>
<td>Stanford</td>
<td>59%</td>
<td>32%</td>
<td>41%</td>
</tr>
<tr>
<td>Stanford-twitter</td>
<td>54%</td>
<td>45%</td>
<td>49%</td>
</tr>
<tr>
<td>Ritter</td>
<td>73%</td>
<td>49%</td>
<td>59%</td>
</tr>
</tbody>
</table>

Table 1: Whole-pipeline named entity recognition performance in the Ritter dataset

Then we utilize OpenCalais\(^1\), due to state-of-the-art semantic functionality and a higher rate limit (50000 per day), which allows for providing unique URIs for the topics so that the meaning of such concepts is well defined.

Scoring User Interests Once the user genuine interest are identified, we then will score them to find the weight of topics of interest of user across different entities. The process is very important as the scores of user’s interests will be utilized to decide the appropriate interests distribution for the user. We employ a frequency-based and confidence-based scoring mechanism. We again use OpenCalais to perform the task.

Definition 2 (Tweet Score). The topic score of each tweet $T$, posted by given a user $u$, can be measured by frequency and confidence for a entity $e$, where a tweet $t$ is represented via a set of entities $E$.

$$S(u, T) = \sum_{i=1}^{n} f_e, c_e, \quad \text{where} \quad f_e, c_e, \in F_E, C_E$$  \hfill (2)

Here, $F_E$ is a set of frequency with $E$ and $C_E$ is a set of confidence with entity $e$ for the given user $u$.

\(^1\)http://www.opencalais.com/
Finally, each of user’s tweets with the weight score are added as leaf nodes to the THT by linking to their appropriate categories.

4 Experiments

4.1 Preparing the datasets

To evaluate our proposed method on real-world data, we collect Twitter information streams of 5,140 randomly selected users over a 3 month period from March to May, 2014. Due to the restriction enforced by Twitter API, for users who have more than 3200 tweets, we could retrieve only 3200 per user. Finally, we obtain more than 10 million tweets published by these people. The results of our measurements show that the number of Twitter messages posted per user follows a power-law distribution. As far as we know, it is ample dataset of sufficient size to perform our analysis.

As we are interested in analyzing a long-term temporal characteristics of user interests, we first select those users which produced more than 1,000 tweets (no spam) that guarantee activity of user and a long-term post behavior. We process each tweet via the semantic enrichment component of our user interests model framework to identify topics and concepts mentioned in the tweets. Table 2 shows finally the selected user statistics, the volume of tweets and number of entities identified in the tweets.

<table>
<thead>
<tr>
<th>Users</th>
<th>Tweets</th>
<th>Concepts</th>
<th>Distinct Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>23</td>
<td>56678</td>
<td>34088</td>
</tr>
<tr>
<td>Average</td>
<td>2464</td>
<td>1482</td>
<td>474</td>
</tr>
</tbody>
</table>

Table 2: Counts and statistics for the tweet dataset used for experimentation

4.2 Evaluation methodology

Our main goal is to analyze and compare how the topics of interest of Twitter user change over time. So we select cosine similarity to evaluate the differences of current-based interests and past-based interests.

As we described in the previous section, tweets are classified into a predefined set of topic categories in THT, \( C = \{c_1, c_2, \ldots, c_n\} \), including ”education”, ”finance”, and ”science”, etc. In our tweets analysis, we computed the topics distribution of interest of user over the set of topic categories for individual user as well as all of users in the sample dataset. First, we select each month of last year (Jan 1th - Dec 31st, 2013) as a time unit. Then, for each user \( u \), we computed the distribution of her interests in every month \( m \), \( D(u, m) \), represented as a vector over the set of topic categories:

\[
D(u, m) = \left( \frac{N_{c_1}}{N}, \frac{N_{c_2}}{N}, \ldots, \frac{N_{c_n}}{N} \right) \quad \text{where} \quad N = \sum_i N_{c_i} \tag{3}
\]

\( N_{c_i} \) is the number of tweets classified into category \( c_i \) published by user \( u \) in a given month \( m \). \( N \) is the total number of tweets posted by the user in the time period. Thus, \( D(u,m) \) represent the proportion of tweets made by the user publishing about each topic category and also reflect the importance of each of interests.
Definition 3 (Interest Similarity). Given a user $u$ interests vector $D(u, m_1)$ at time $m_1$ and $D(u, m_2)$ at time $m_2 (m_1 < m_2)$, which are represented via topics distribution of interests using the same vector representation that is used for a given user interests, the similarity algorithm ranks user interests change degree according to their cosine similarity.

\[
sim_I(D(u, m_1), D(u, m_2)) = \frac{D(u, m_1) \cdot D(u, m_2)}{\|D(u, m_1)\| \|D(u, m_2)\|}
\] (4)

Here, if the value of $\sim_I$ is low, it indicates the interests of same user has a large change over time. It is possible reason that the user change his job or admit to university. On the contrary, the user has a steady habit of interests.

To further examine the performance of those models in inferring user interests, we also select two representative the ranking system of user interests (detailed in Section 5.3) and compute the top interests for them by using the three models.

5 Results

5.1 User Interests Distribution

Twitter, one of the most popularity of social network, relies on the idea that anyone can contribute and discuss her opinion about all kind of topics. Therefore, we employ our proposed THT to generate topic-based user interests profile. The profile features 23 different dimensions that correspond to broad topics like technology or education.

![Figure 3: The screenshot of the pie chart of users interests in the dataset](image)

Figure 3 gives the pie chart visualization of topic-based interests on our constructed dataset, we can see about which topics has been published most in Twitter. The result shows that 13.5% of tweets are concerned with technology. This observation is in line with the reports we got from categorization of Twitter users\(^2\). Through the pie chart, we can moreover compare this interest with interests for other topics. For example, we observe that 9.0% of users is interested in "health" more than interested in education or shopping.

![Figure 4: The distribution of topics of interests of users](image)

Figure 4 plots the topic distribution of interests discussed by 3 users who are randomly selected. For the clarity of the figure, all of topic categories are shown in the figure. We can see

\(^2\)http://www.beevolve.com/twitter-statistics/
that the distribution of each of user interests has a greater difference. In general, as humans, we have limited knowledge and limited time and limit attention. Thus, the user interests can be divided into primary topics of interest and secondary topics of interest. Intuitively, the primary interest of user hold stability over time; the secondary interest, however, is more likely to keep up with hot topics or events in the moment. We will explore this issue in the next section.

5.2 Change of User Interests over Time

If a user’s interests are stable, her interests distributions in each month should be stability over time. Particularly, we are interested in hierarchy of needs of user interests. To accomplish the task, for each Twitter user, we compared her interests distribution of the most recent month to her interests distributions of all the previous months (Jan/1/2013 - Dec/31/2013). To validate our hypothesis mentioned the above section, we randomly selected 10 users from user set and applied different strategies to create semantic user interests from their Twitter messages.

Figure 5: Similarity of user interests between current month and previous months

Figure 6: Temporal evolution of top 5 topics of interest of a Twitter’s user

Figure 5 depicts the three different type of similarity of the user current interests with the interests of the same users created based on Twitter activities performed in a certain month in the past. We can see the similarity of current interests with the corresponding interests generated at the beginning of our observation period is the lowest while the similarity of current interests with interests computed last month is the highest. The comparison demonstrates fresh interest profiles better suit user current demands. Furthermore, Entity-based interests exhibit the weakest changes over time as the average similarity to the current interest is constantly lower than for the topic-based and concept-based interests.

We also explore whether user’s interests have hierarchy or not. The result is illustrated in Figure 6. Due to the limit space, we only show top 5 topics of interests of a user change over time, other users haves the similar case. From the figure, we conclude that the user interests can be divided into primary interest (corresponding to the upper half part) and secondary interest (corresponding to the lower half part). The primary interest of user is an original inherent preferences, such as programmers like efficient algorithm or lawyers like debate, and thus keep higher ratio and hold stability in a long-term period. However, the secondary interest is an amateur demand and more likely to closely follow the tracks of hot topics or events in the moment. This is a good evidence that can be answered the mentioned problem in Section 5.1.
5.3 Comparative Evaluation

The state-of-the-art web service of user’s interests that has been profiled is two systems which called *Who likes what?* [3] and TUMS [16]. The first system generates general interests and niche interests as user interests from received tweets. The second system generates semantic user interests from tweets. Table 3 shows the declared interests for three users (as given in their description) and the top topics of interest inferred by the three models.

<table>
<thead>
<tr>
<th>Self-description of user</th>
<th>Top topics of interest, inferred by different models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Who likes what?</td>
</tr>
<tr>
<td>user A: develop extra income without any financial risk</td>
<td>marketing, business, social media, world, internet</td>
</tr>
<tr>
<td>user B: passionate parent, education nonprofit exec, mom congress delegate, foodie/techie’s wife</td>
<td>education, news, media, parenting</td>
</tr>
<tr>
<td>user C: data recovery, computer forensics, investigations, data conversion, disaster recovery</td>
<td>data recovery, business, social media, marketing, bloggers</td>
</tr>
</tbody>
</table>

Table 3: Examples of top topics of interests of user inferred by the three models, for some Twitter users who profile their interests in their self-description.

In Table 3 we summarize the user interests obtained by each of the above models. As we can see the the first two model mostly accurately captured their broad interests, yet the interests were sometimes too general. Our model exhibit a finer resolution of topics of interest of Twitter user on our proposed Topic Hierarchy Tree. The results of our analysis can be provided more precise directions for personalization recommended system to improve performance.

6 Conclusions and Future Work

In this paper we developed a user interests modeling framework for Twitter to analyze time-sensitive characteristics of user interests. The temporal analysis thus revealed two important observations. First, user interests change over time: fresh interest profiles seem to better suit user current demands. Second, user interests have the characteristic of hierarchy of needs.

In future work we will further research the problem of multi-sources user interests using semantic technologies. The same user is more possible profile overall interests via multiple social networks. Therefore we plan to explore whether multi-sources user interests can further leverage personalization quality for recommender system.

Acknowledgement

This work was supported by National Key Technology R&D Program(No.2012BAH46B03), and the Strategic Leading Science and Technology Projects of Chinese Academy of Sciences(No.XDA06030200).
References


