Predicting Users' Future Interests on Twitter

Fattane Zarrinkalam[1,2(✉)], Hossein Fani[1,3], Ebrahim Bagheri[1], and Mohsen Kahani[2]

[1] Laboratory for Systems, Software and Semantics (LS3), Ryerson University, Toronto, Canada
[2] Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
fattane.zarrinkalam@gamil.com
[3] Faculty of Computer Science, University of New Brunswick, Fredericton, Canada

Abstract. In this paper, we address the problem of predicting future interests of users with regards to a set of *unobserved* topics in microblogging services which enables forward planning based on potential future interests. Existing works in the literature that operate based on a known interest space cannot be directly applied to solve this problem. Such methods require at least a minimum user interaction with the topic to perform prediction. To tackle this problem, we integrate the semantic information derived from the Wikipedia category structure and the temporal evolution of user's interests into our prediction model. More specifically, to capture the temporal behaviour of the topics and user's interests, we consider discrete intervals and build user's topic profile in each time interval separately. Then, we generalize users' interests that have been observed over several time intervals by transferring them over the Wikipedia category structure. Our approach not only allows us to generalize users' interests but also enables us to transfer users' interests across different time intervals that do not necessarily have the same set of topics. Our experiments illustrate the superiority of our model compared to the state of the art.

1 Introduction

Techniques for the identification and modeling of user interests based on users' social presence have received much attention in the recent years [2,10]. Researchers have already explored ways in which user interests can be modeled in social networks with special attention being given to Twitter. Existing works often provide a view of users' interests with regards to a set of core themes. For instance, some works have expressed users' interests in terms of bag of words, Wikipedia entries or in relation to the current active topics on the social network.

While approaching the problem from different technical perspectives, most of the existing works on social networks focus on modeling users' current interests and little work has been done on the prediction of users' potential future

J.M. Jose et al. (Eds.): ECIR 2017, LNCS 10193, pp. 464–476, 2017.
DOI: 10.1007/978-3-319-56608-5_36

interests. In all these works, the interest space is assumed to be known *a priori*; therefore, various models of collaborative filtering and link prediction that require a known interest space can effectively be employed [3,23].

Our work in this paper aims to extend the state of the art by predicting users' interests with regards to future *unobserved* topics. In other words, our objective is to provide a solution for performing *what-if* analysis over potential future topics. For instance, we are interested in determining whether a given user would be interested in following the news about the release of a new mobile operating system that would compete with iOS. Our work will enable forward planning based on potential future interests. Given the focus of our work on unobserved topics, existing works in the literature that operate based on a known interest space cannot be directly applied to it. Those techniques would require at least some minimum user interactions [4].

To address the above problem statement, in this paper, we propose a prediction framework to integrate semantic information from knowledge bases such as Wikipedia and temporal evolution of each individual user's interests to predict user's future interests. Knowledge infused prediction algorithms have gained significant attention due to their competitive performance and ability to overcome the cold start problem [14,18]. However utilizing knowledge bases for improving user interest prediction methods in microblogging services is largely unexplored. Our prediction model is based on the intuition that, although it is possible that the topics of interest to the users dramatically change over time as influenced by real-world trends [1], users tend to incline towards topics and trends that are semantically or conceptually similar to a set of core interests. Therefore, in order to be able to achieve predictability, one would need to generalize each individual user's interests over several time intervals to gain a good insight into the user's overall mindset. To this end, we generalize users' interests that have been observed over several time intervals by transferring them onto the Wikipedia category structure. Generally, our approach utilizes the Wikipedia category structure to model high level user interests and takes the temporal evolution of user's interests into account in order to predict user's future interests. The key contributions of this paper are as follows:

- We propose a model that transfer user's interests from different time intervals onto the Wikipedia's category structure. In this process, we model high-level interests of users such that the evolution of user's interests over topics is captured.
- We illustrate how semantic information derived from the Wikipedia knowledge base as well as temporal information can be integrated in our model to predict user's interests with regards to unobserved topics of the future in Twitter.
- We perform experimentation to illustrate the impact of considering Wikipedia categories on the accuracy of predicting the future interests of users on Twitter. The experimental results demonstrate the superiority of our model compared to the state of the art methods which tackle cold item problem.

The rest of the paper is organized as follows: In Sect. 2 we describe the related work. Sections 3 and 4 are dedicated to the problem definition and the presen-

tation of the details of our proposed approach. Section 5 presents the details of our experimental work. Finally, Sect. 6 concludes the paper.

2 Related Work

There is a rich line of research on user interest detection from social networks through the analysis of user generated textual content. To represent user interests, such works either use *Bag of Words*, *Topic Modeling* or *Bag of Concepts* approach. Since the *Bag of Words* [20] and *Topic Modeling* [19] approaches focus on terms without considering their semantics and the relationship between them, they do not necessarily utilize the underlying semantics of textual content. Furthermore, these approaches may not perform so well on short, noisy and informal texts like Twitter posts [6]. To address these issues, the *Bag of Concepts* approach utilizes external knowledge bases to enrich the representation of short textual content and model user interests through semantic entities (concepts) linked to external knowledge bases such as DBpedia. Since these knowledge bases represent entities and their relationships, they provide a way of inferring underlying semantics of content [13].

While existing work on microblogging services mainly focus on extracting users' current interests, little work has been done on predicting users' future interests. Bao et al. [3] have proposed a temporal and social probabilistic matrix factorization model that utilize users' sequential interest matrices at different time intervals and the users' friendships matrix to predict future users' interest in microblogging services. Their work is very similar to ours in a sense that we both try to predict future user interests in microblogging services by taking into account the temporal evolution of user interests. However, they are limited by the fact that they assume the topic set of the future to be known *a priori* and composed only of the set of topics that have been observed in the past. Therefore, they cannot predict user interests with regard to new topics since these topics have never received any feedbacks from users in the past.

Given users' interests change over time, temporal aspects have been widely used for the conventional recommendations and user modeling in online social networks [21]. Many researchers have focused on applying time decay functions over historical user generated content [8]. Based on time decay functions, the weight of each interest is calculated depending on its age. Recently, Piao and Breslin [15] have studied the effectiveness of different time decay functions for incorporating dynamics of user interests in the context of personalized link recommendations on Twitter. They have shown that using decay functions to build users' long-term profiles results in noticeable improvement in the quality of recommendations compared to user profiles without considering any decay of user interests. There is another line of related works that utilize knowledge base information to overcome the cold start problem in traditional algorithms in the context of recommender systems [11]. For example, Cheekula et al. [5] have proposed a content-based recommendation method that utilizes hierarchical user interests over Wikipedia category hierarchy to identify relevant entities. Their

work is similar to ours in a sense that both model high-level interests of users over the Wikipedia category graph. However, they overlook the evolution of user's interests over time. Further, our work focuses on predicting user's interests over unobserved topics in the future as opposed to entity recommendation.

3 Preliminaries

3.1 User Interest Profile

In our work, we model users' interests in relation to the active topics of the social network. A topic z has traditionally been defined as a semantically coherent theme which has received substantial attention from the users.

Let t be a specified time interval, given $\mathbb{Z}^t = \{z_1^t, z_2^t, \ldots, z_K^t\}$ be K active topics in t, for each user $u \in \mathbb{U}$, we define her topic profile in time interval t, $TP^t(u)$, which is the distribution of u's interests over $\mathbb{Z}^t$, as follows:

Definition 1 ***(Topic Profile).*** *The topic profile of user $u \in \mathbb{U}$ in time interval t, with respect to $\mathbb{Z}^t$, denoted by $TP^t(u)$, is represented by a vector of weights over the K topics, i.e., $(f_u^t(z_1^t), \ldots, f_u^t(z_K^t))$, where $f_u^t(z_k^t)$ denotes the degree of u's interest in topic $z_k^t \in \mathbb{Z}^t$. A user topic profile is normalized so that the sum of all weights in a profile equals to 1.*

It should be noted that topic and user interest detection methods from microblogging services have already been well 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 $\mathbb{Z}^t$ and $TP^t(u)$.

3.2 Problem Definition

The objective of our work is to answer *what-if* questions by predicting user interests with regards to potentially trending topics of the future. To achieve this goal, we rely on temporal and historical user interest information in order to predict how users would react to future topics. Recent studies have already shown that trending topics on social networks can rapidly change in reaction to real world events and therefore, the set of topics might significantly change between different time intervals [1]. Therefore, to express the temporal dynamics of topics and user interests, we divide the users' historical data into L discrete time intervals $1 \leq t \leq L$ and extract L topic sets $\mathbb{Z}^1, \mathbb{Z}^2, \ldots, \mathbb{Z}^L$, in these time intervals using the microposts which are published in each time interval separately. More specifically, for each time interval $t : 1 \leq t \leq L$, we first extract active topics in that time interval $\mathbb{Z}^t$, and then for each user $u \in \mathbb{U}$, we build her topic profile in time interval t, $TP^t(u)$, as a result of which each user will have L user profiles, one for each of the time intervals. Informally speaking, our objective is to exploit the L historical topic profiles of a user u, to predict the user's inclination towards the topics of time interval $L + 1$.

Definition 2 ***(Future Topic Profile).*** *Given the topic profiles for each user u in each time interval of the historical data, $TP^1(u), \ldots, TP^L(u)$, and a set of topics in time interval $L+1$, $\mathbb{Z}^{L+1}$, which might not have been observed in the previous time intervals, we aim to predict $\widehat{TP}^{L+1}(u)$, the future topic profile of user u towards $\mathbb{Z}^{L+1}$.*

To address the challenge defined in Definition 2, we divide this problem into two subproblems: *historical user topic profile extraction* and *future interest prediction*, in which the output of the first subproblem becomes the input of the second one.

4 Proposed Approach

In this section, we first introduce our method to extract historical topic profile of users and then we describe our prediction model to predict future interests of users.

4.1 Historical User Topic Profile Extraction

As explained earlier, our work relies on each user's topic profiles within the past L intervals. Each user topic profile in a given time interval t is a distribution over the active topics in that time interval $\mathbb{Z}^t$, which is not necessarily the same as the topics in the previous or next time intervals. In order to extract $TP^t(u)$, the user topic profile for each user u in each time interval of the historical data, $1 \leq t \leq L$, we employ the LDA topic modeling approach.

Considering $\mathbb{M}^t$, the set of microposts as a text corpus published in time interval t, it is possible to extract topics $\mathbb{Z}^t$ using topic modeling methods. As proposed in [16], to obtain better topics from microblogging services without modifying the standard topic modeling methods, we enrich each micropost m from our corpus $\mathbb{M}^t$ by using an existing semantic annotator and employ the extracted entities, which can lead to the reduction of noisy content within the topic detection process. Therefore, in our work, each micropost is considered as a set of one or more semantic entities that collectively denote the underlying semantics of the microposts. Therefore, we view a topic, defined in Definition 3, as a distribution over Wikipedia entities.

Definition 3 ***(Topic).*** *Let $\mathbb{M}^t$ be a corpus of microposts published in time interval t and $\mathbb{E} = \{e_1, e_2, \ldots, e_{|\mathbb{E}|}\}$ be the vocabulary of Wikipedia entities, an active topic in time interval t, z^t, is defined to be a vector of weights, i.e., $(g_z^t(e_1), \ldots, g_z^t(e_{|\mathbb{E}|}))$, where $g_z^t(e_i)$ shows the participation score of term $e_i \in \mathbb{E}$ in forming topic z^t. Collectively, $\mathbb{Z}^t = \{z_1^t, z_2^t, \ldots, z_K^t\}$ denotes a set of K topics extracted from $\mathbb{M}^t$.*

To extract the topics from microposts using LDA, documents should naturally correspond to microposts. However, since our goal is to understand the topics that each user u is interested in rather than the topic that each single

micropost is about, similar to previous works in the literature [17], we aggregate the published or retweeted microposts of a user u in time interval t, i.e., $\mathbb{M}_u^t$, into a single document. LDA has two parameters to be inferred from the corpus of documents: document-topic distributions θ, and the K topic-term distributions ϕ. Given that each document corresponds to a user u and Wikipedia entities $\mathbb{E}$ as the vocabulary of terms, by applying LDA over the microposts $\mathbb{M}^t$, the results produce the following two artifacts:

- K topic-entity distributions, where each topic entity distribution associated with a topic $z^t \in \mathbb{Z}^t$ represents active topics in $\mathbb{M}^t$, i.e., $(g_z^t(e_1), \ldots, g_z^t(e_{|\mathbb{E}|}))$
- $|\mathbb{U}|$ user-topic distributions, where each user-topic distribution associated with a user u, represents the topic profile of user u in time interval t, i.e., $TP^t(u) = (f_u^t(z_1^t), \ldots, f_u^t(z_K^t))$.

Now, given a corpus of microposts $\mathbb{M}$, we will break it down into L intervals and perform the above process separately on each of the intervals. This will produce $TP^1(u), \ldots, TP^L(u)$ for every user u in our user set, which is the required input for our future user interest prediction problem defined in Definition 2.

4.2 Future Interest Prediction

Given $TP^1(u), TP^2(u), \ldots, TP^L(u)$, our goal is to predict potential interests of each user u over $\mathbb{Z}^{L+1}$. It is important to point out that since $L+1$ is in the future, the topics $\mathbb{Z}^{L+1}$ have not yet been observed. Therefore, our work aims to answer important *what-if* questions in that it is able to predict how the users react to a given set of topics. This allows one to perform future planning by studying how users will react if certain topics emerge in the future. Our prediction model is based on the intuition that while user interests might change over time, they tend to revolve around some fundamental issues. More specifically, although user interests are driven by the shifts and changes in real world events and trends [1], they tend towards topics and trends that are semantically or conceptually similar. For this reason, we generalize users' interests that have been observed over several time intervals by transferring them over the Wikipedia category structure. This approach will not only allow us to generalize users' interests but also enables us to transfer users' interests across different time intervals that do not necessarily have the same set of topics.

Based on the above intuition, formally, for each user u, given the topic profiles of the user u in each time interval t, $TP^t(u)$, we utilize Wikipedia category structure to build a category profile for user u in each time interval t, denoted as $CP^t(u)$.

Definition 4 ***(Category Profile).*** *The category profile of user $u \in \mathbb{U}$ in time interval t toward Wikipedia categories $\mathbb{C} = \{c_1, c_2, \ldots, c_{|\mathbb{C}|}\}$, called $CP^t(u)$, is represented by a vector of weights, i.e., $(h_u^t(c_1), \ldots, h_u^t(c_{|\mathbb{C}|}))$, where $h_u^t(c)$ denotes the degree of u's interest in category $c \in \mathbb{C}$ at time interval t. A user category profile is normalized so that the sum of all weights in a profile equals to 1.*

Now, based on the Category Profiles of each user derived from the past L consecutive time intervals, $CP^1(u), \ldots, CP^L(u)$, we apply our model to predict $\widehat{TP}^{L+1}(u)$.

Category Profile Identification. In this section, we aim at utilizing the Wikipedia category structure to generalize the topic-based representation of user interests to category-based representation. To do so, there are two possible approaches through which we build the category profile of a user u at time interval t, $CP^t(u)$, given her topic profile $TP^t(u)$: *(1) attribution*, and *(2) hierarchical* approach.

In the *attribution* approach, for each user u, only those categories that are directly associated with the constituent entities of the user's topics of interest are considered as categories of interest. We essentially map $TP^t(u) = (f_u^t(z_1^t), \ldots, f_u^t(z_K^t))$ to $CP^t(u) = (h_u^t(c_1), \ldots, h_u^t(c_{|\mathbb{C}|}))$ as follows:

$$h_u^t(c) = \sum_{i=1}^{K} f_u^t(z_i^t) \times \Phi(z_i^t, c) \tag{1}$$

where $\Phi(z, c)$ denotes the degree of relatedness of topic $z^t = (g_z^t(e_1), \ldots, g_z^t(e_{|\mathbb{E}|}))$ to category $c \in \mathbb{C}$ and is calculated based on Eq. 2.

$$\Phi(z^t, c) = \sum_{i=1}^{|\mathbb{E}|} g_z^t(e_i) \times \delta_c(e_i) \tag{2}$$

Here, $\delta_c(e)$ is set to 1 if entity e is a Wikipedia page that belongs to the Wikipedia category c, otherwise it is zero and $g_z^t(e)$ is the distribution value of entity e in topic z^t, produced by applying LDA over $\mathbb{M}^t$ as described in Sect. 4.1. It is important to note that the reason why we can calculate the relatedness of each topic to each category is that we view each topic as a distribution over Wikipedia entities and in Wikipedia, each entry is already associated with one or more categories.

In the *hierarchical* approach, we assume that when a user is interested in a certain category, she might also be interested in broader related categories. Based on this, in the hierarchical approach, we first infer the broadly related categories of user interests by exploiting the hierarchy of the Wikipedia category structure. A major challenge in utilizing Wikipedia category structure as a hierarchy is that, it is a cyclic graph instead of a strict hierarchy [9]. Therefore, as a preprocess in the hierarchical approach, we transform the Wikipedia category structure into a hierarchy by adopting the approach proposed in [9]. The output of this process is a Wikipedia Category Hierarchy (WCH), a directed acyclic graph whose nodes are the Wikipedia categories $\mathbb{C}$ with an edge from $c_i \in \mathbb{C}$ to $c_j \in \mathbb{C}$ whenever c_i is a subcategory of c_j.

For a user u, given $TP^t(u) = (f_u^t(z_1^t), \ldots, f_u^t(z_K^t))$ and Wikipedia Category Hierarchy WCH as input, we infer the hierarchical interests of user u in time interval t, represented in the form of a category hierarchy. To do so, for each

topic z_i^t, we first assign an initial score of $f_u^t(z_i^t) \times \Phi(z_i^t, c)$ to every category node $c \in \mathbb{C}$ similar to what is done in the attribution approach. Then, the score of each category node with a $score(c) > 0$ is propagated up the hierarchy as far as the root using a Spreading Activation function to calculate the new score of each node. We adopt the 'Bell Log' function as our spreading activation function as described in [9].

Now, given topic profiles of a user u in L consecutive time intervals of the historical data, i.e., $TP^1(u), \ldots, TP^L(u)$, we perform the above process separately on each of the intervals. This will produce $CP^1(u), \ldots, CP^L(u)$ for every user $u \in \mathbb{U}$, which is the input of our method described in the next section to predict $\widehat{TP}^{L+1}(u)$.

Interest Prediction. Given $CP^1(u), \ldots, CP^L(u)$, our first step to predict $\widehat{TP}^{L+1}(u)$ is calculating $CP^{L+1}(u)$. As already discussed in the literature, users' current interests are driven by their past interests, interactions and behavior where distant history has a lesser influence on the current interests compared to more recent events and activities [15]. Based on this observation, we employ a decay function in order to soften the impact of distant experiences on the users' future interests. We choose the exponential decay function which can describe this influence effectively [8]. More formally, we calculate the category profile of user u in time interval $L+1$, $CP^{T+1}(u) = (h_u^{L+1}(c_1), \ldots, h_u^{L+1}(c_{|\mathbb{C}|}))$, as follows:

$$h_u^{L+1}(c) = \sum_{t=1}^{L} exp(-\frac{L-t}{\alpha}) h_u^t(c) \tag{3}$$

where the value of $\alpha > 0$ presents the kernel parameter, and the value of L shows the number of time intervals that the historical data is divided to. In our experiments, we choose α as the length of each time interval t [12].

Given the high-level interests of user u in time interval $L+1$ represented over Wikipedia categories, $CP^{L+1}(u)$, and a set of unobserved topics (*what-if* subjects) for time interval $L+1$, $\mathbb{Z}^{L+1}$, we are interested in predicting a topic profile for user u, $\widehat{TP}^{L+1}(u) = (\hat{f}_u^{\,L+1}(z_1^{L+1}), \ldots, \hat{f}_u^{\,L+1}(z_K^{L+1}))$. We calculate $\hat{f}_u^{\,L+1}(z_i^{L+1})$ as follows:

$$\hat{f}_u^{\,L+1}(z_i^{L+1}) = \sum_{j=1}^{\mathbb{C}} \Phi(z_i^{L+1}, c_j) \times h_u^{L+1}(c_j) \tag{4}$$

where $\Phi(z, c)$ calculates the relatedness of topic z to category c based on Eq. 2.

5 Experiments

5.1 Dataset and Experimental Setup

In our experiments, we use an available Twitter dataset collected and published by Abel et al. [2]. It consists of approximately 3M tweets posted by 135,731

unique users. We annotated the text of each tweet with Wikipedia entities using the TAGME RESTful API[1], which resulted in 350,731 unique entities. We divide our dataset into $L+1$ fixed time intervals. The first L time intervals serve as our training data and the last is employed for testing. To prepare Wikipedia category graph, we downloaded the freely available English version of DBpedia, which is extracted from Wikipedia dumps dating from October 2015. This dataset consists of 968,350 categories with 2,225,459 subcategory relations between them. We preprocessed the Wikipedia category hierarchy as suggested in [9]. The outcome of this process is a hierarchy with a height of 20 and 824,033 categories with 1,506,292 links among them.

5.2 Evaluation Methodology and Metrics

Given the outputs of LDA over $L+1$ time intervals of our dataset, we consider the first L extracted topic profiles of each user u, $TP^1(u), TP^2(u), \ldots, TP^L(u)$, as her historical interests for training and $TP^{L+1}(u)$ as the golden truth of her interests for testing.

To evaluate $\widehat{TP}^{L+1}(u)$, we choose two popular metrics for evaluating the '*accuracy of predictions*': Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). A lower MAE or RMSE scores indicates more accurate prediction results. Further we calculate the Normalized Discounted Cumulative Gain (nDCG) as a well-known metric for evaluating the '*ranking quality*' of the results.

5.3 Comparison Methods

Our goal is to predict the degree of user interests over topics that emerge in the future, which have not been observed in the past. Among different recommendation strategies, collaborative filtering methods cannot recommend new items since these items have never received any user's feedbacks in the past. To tackle the cold item problem, content-based and hybrid approaches that incorporate item content are recommended [4]. Thus, we consider the following comparison methods:

SCRS (Semantic Content-based Recommender System) [14] extracts item features from Wikipedia to compute the semantic similarity of two items. The adoption of this approach in our context would need us to consider each topic of interest as an item and the constituent Wikipedia entities of a topic as its content. Then, we predict $\widehat{TP}^{L+1}(u) = (\hat{f}_u^{L+1}(z_1^{L+1}), \ldots, \hat{f}_u^{L+1}(z_K^{L+1}))$ as follows:

$$\hat{f}_u^{L+1}(z_i^{L+1}) = \frac{1}{K \times L} \sum_{t=1}^{L} \sum_{j=1}^{K} f_u^t(z_j^t) \times S(z_i^{L+1}, z_j^t) \tag{5}$$

where $S(z_1, z_2)$ denotes the similarity of two topics calculated by the cosine similarity of their respective entity weight distribution vectors defined in Definition 3.

[1] http://tagme.di.unipi.it/.

ACMF (Attribute Coupled Matrix Factorization) [22] is a hybrid approach that incorporates item-attribute information (item content) into the matrix factorization model to cope with the cold item problem. In our work, the items are the topics of all time intervals, i.e., $\mathbb{Z} = \bigcup_{1 \leq t \leq L+1} \mathbb{Z}^t$. Accordingly, the item relationship regularization term is adopted as follows:

$$\frac{\beta}{2} \sum_{i=1}^{|\mathbb{Z}|} \sum_{j=1}^{|\mathbb{Z}|} S(z_i, z_j) ||q_i - q_j||_F^2 \tag{6}$$

where β is the regularization parameter to control the effect of the item (topic)-attribute information, $S(z_1, z_2)$ is the similarity between topics z_i and $z_j \in \mathbb{Z}$, as described for Eq. 5. Further, q is the topic latent feature vector, and $||.||_F^2$ is the Frobenius norm.

Attribution (Attribution-based future user interest prediction) is a variant of our proposed approach which uses the attribution method as described in Sect. 4.2 to build the category profile of a user.

Hierarchical (Hierarchical-based future user interest prediction) is a variant of our proposed approach which uses the Wikipedia Category Hierarchy as described in Sect. 4.2 to build category profile of a user.

5.4 Results and Discussion

In order to ensure that our experiments are generalizable and not impacted by the effect of parameter setting, we explore a range of values for the two possible variables that can affect the performance of our work, i.e., the length of the time intervals and the number of topics. We perform the evaluations for different lengths of time interval: 1, 3 and 7 days and for varying number of topics ranging from 20 to 50. We present the quality of the prediction results in Fig. 1 where we can observe that the two variants of our proposed approach, i.e., Attribution and Hierarchical methods, outperform SCRS and ACMF in terms of both MAE and RMSE. This observation confirms that utilizing Wikipedia category structure enables us to model user's high level interests more accurately and consequently can lead to improve the quality of user interest prediction with regards to new topics of the future. It is worth noting that this achievement is consistent in all different time interval sizes and the number of topics.

Figure 1 additionally shows that our method (Attribution) outperforms the other comparison methods in terms of the ranking metric (nDCG). This is an important observation when it is considered collectively with the results obtained from MAE and RMSE. It points to the fact that the Attribution method not only provides an accurate estimation of the degree of interest but is also able to accurately predict the ranking of user interests, which shows that we can estimate the *preference order* between user interests as well as the degree of difference between these interests for every given user. Now, when considering the other baseline methods, it is interesting to see that while SCRS performs the worst among the various methods in terms of MAE and RMSE, it produces accurate

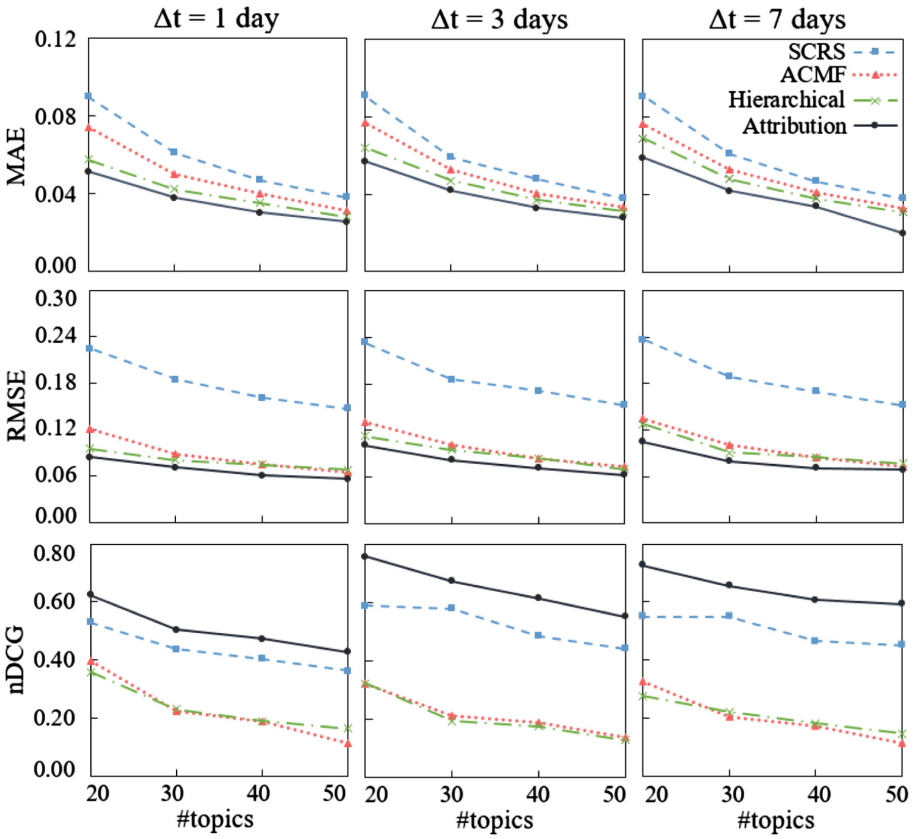

Fig. 1. Evaluation results in terms of MAE, RMSE and nDCG.

rankings. This can potentially be explained by the fact that while SCRS is not able to accurately predict the degree of user interests, it is able to estimate the preference order between the user interests. However, our proposed Attribution approach is still the best performing method in all three measures.

By comparing Attribution and Hierarchical variants of our proposed approach, one can observe that the Attribution method provides better results. Both methods model user high-level interests over Wikipedia categories. The difference is that, in the Attribution approach, only those categories that are directly associated with the constituent entities of the user's topics of interest are considered as categories of interest. However, in the Hierarchical approach, broadly related categories of user interests are also considered by applying a spreading activation function over the hierarchy of the Wikipedia category structure. Here we adopt the Bell-log activation function proposed in [9] for this purpose. We speculate the probable cause for the poor performance of the Hierarchical approach compared to Attribution approach is the Bell-log activation function. On the one hand, Bell-log activation function spreads all the scores from the

leaves up to the root of the hierarchy in a way that broader categories receive higher scores. On the other hand, higher categories are usually common among majority of users' category profiles. In the prediction step, it may happen that a topic can belong to this very broad category. Hence, this topic will be predicted as a topic of interest for almost all users which leads to the above mentioned poor accuracy. We believe discrete time state space models [7] may alleviate the inappropriate score assignments by the Bell-log. Such models set category score based on a convex combination of its predecessors and successors. This will be another area for our future investigation.

Now, among the baselines and as shown in Fig. 1, ACMF, which is a hybrid recommender system that combines collaborative filtering and topic content, can achieve more accurate results in terms of MAE/RMSE in comparison with SCRS, which is solely based on topic content. This could indicate that incorporating user interests of other users might improve the accuracy of user interest predictions. Based on this observation, it seems promising to investigate collaborative extensions of our proposed approach as future work.

6 Conclusions

In this paper, we address the problem of predicting future interests of users with regards to a set of *unobserved* topics (*what-if* subjects) on Twitter. Our model is based on the intuition that while user interests might change over time in reaction to real world events, they tend to revolve around some fundamental issues that can be seen as the user's mindset. To capture the temporal behaviour of the topics and user's interests, we consider discrete time intervals and build user's topic profile in each time interval as the user's historical topic profiles. Then, we generalize each individual user's topic profile as we move through time from the oldest to the most recent interval to infer the user category profile using the Wikipedia category structure. Given a user category profile, we predict the degree of interest of a user to each unobserved topic based on the relatedness of each topic to the inferred category profile. Our experiments illustrate the superiority of our model compared to the state of the art.

References

1. Abel, F., Gao, Q., Houben, G., Tao, K.: Analyzing temporal dynamics in Twitter profiles for personalized recommendations in the social web. In: WebSci 2011, pp. 2:1–2:8 (2011)
2. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Analyzing user modeling on Twitter for personalized news recommendations. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) UMAP 2011. LNCS, vol. 6787, pp. 1–12. Springer, Heidelberg (2011). doi:10.1007/978-3-642-22362-4_1
3. Bao, H., Li, Q., Liao, S.S., Song, S., Gao, H.: A new temporal and social pmf-based method to predict users' interests in micro-blogging. Decis. Support Syst. **55**(3), 698–709 (2013)

4. Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A.: Recommender systems survey. Knowl.-Based Syst. **46**, 109–132 (2013)
5. Cheekula, S.K., Kapanipathi, P., Doran, D., Jain, P., Sheth, A.P.: Entity recommendations using hierarchical knowledge bases. In: ESWC 2015 (2015)
6. Cheng, X., Yan, X., Lan, Y., Guo, J.: BTM: topic modeling over short texts. IEEE Trans. Knowl. Data Eng. **26**(12), 2928–2941 (2014)
7. Dahleh, M., Dahleh, M.A., Verghese, G.: Lectures on dynamic systems and control. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (2013)
8. Ding, Y., Li, X.: Time weight collaborative filtering. In: International Conference on Information and Knowledge Management, pp. 485–492 (2005)
9. Kapanipathi, P., Jain, P., Venkataramani, C., Sheth, A.: User interests identification on Twitter using a hierarchical knowledge base. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8465, pp. 99–113. Springer, Cham (2014). doi:10.1007/978-3-319-07443-6_8
10. Kapanipathi, P., Orlandi, F., Sheth, A.P., Passant, A.: Personalized filtering of the Twitter stream. In: The Second Workshop on Semantic Personalized Information Management: Retrieval and Recommendation 2011, pp. 6–13 (2011)
11. Khrouf, H., Troncy, R.: Hybrid event recommendation using linked data and user diversity. In: RecSys 2013, pp. 185–192 (2013)
12. Li, L., Zheng, L., Yang, F., Li, T.: Modeling and broadening temporal user interest in personalized news recommendation. Expert Syst. Appl. **41**(7), 3168–3177 (2014)
13. Michelson, M., Macskassy,S.A.: Discovering users' topics of interest on Twitter: a first look. In: AND 2010, pp. 73–80 (2010)
14. Noia, T.D., Mirizzi, R., Ostuni, V.C., Romito, D., Zanker, M.: Linked open data to support content-based recommender systems. In: I-SEMANTICS 2012, pp. 1–8 (2012)
15. Piao, G., Breslin, J.G.: Exploring dynamics and semantics of user interests for user modeling on Twitter for link recommendations. In: SEMANTICS 2016, pp. 81–88 (2016)
16. Varga, A., Basave, A.E.C., Rowe, M., Ciravegna, F., He, Y.: Linked knowledge sources for topic classification of microposts: a semantic graph-based approach. J. Web Semant. **26**, 36–57 (2014)
17. Weng, J., Lim, E., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential Twitterers. In: WSDM 2010, pp. 261–270 (2010)
18. Weng, L., Xu, Y., Li, Y., Nayak, R.: Exploiting item taxonomy for solving cold-start problem in recommendation making. In: ICTAI 2008, pp. 113–120 (2008)
19. Xu, Z., Lu, R., Xiang, L., Yang,Q.: Discovering user interest on Twitter with a modified author-topic model. In: WI 2011, pp. 422–429 (2011)
20. Yang, L., Sun, T., Zhang, M., Mei, Q.: WWW 2012, pp. 261–270 (2012)
21. Yin, H., Cui, B., Chen, L., Hu, Z., Zhou, X.: Dynamic user modeling in social media systems. ACM Trans. Inf. Syst. **33**(3), 10:1–10:44 (2015)
22. Yu, Y., Wang, C., Gao, Y.: Attributes coupling based item enhanced matrix factorization technique for recommender systems. CoRR, abs/1405.0770 (2014)
23. Zarrinkalam, F., Fani, H., Bagheri, E., Kahani, M.: Inferring implicit topical interests on Twitter. In: Ferro, N., Crestani, F., Moens, M.-F., Mothe, J., Silvestri, F., Nunzio, G.M., Hauff, C., Silvello, G. (eds.) ECIR 2016. LNCS, vol. 9626, pp. 479–491. Springer, Cham (2016). doi:10.1007/978-3-319-30671-1_35