Information Processing and Management 54 (2018) 339–357

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Mining user interests over active topics on social networks

Fattane Zarrinkalam[a,b], Mohsen Kahani[*,b], Ebrahim Bagheri[a]

[a] *Laboratory for Systems, Software and Semantics (LS^3), Ryerson University, Canada*
[b] *Department of Computer Engineering, Ferdowsi University of Mashhad, Iran*

ARTICLE INFO

Keywords:
User interest detection
Social network analysis
Homophily
Topic relatedness
Semantic annotation

ABSTRACT

Inferring users' interests from their activities on social networks has been an emerging research topic in the recent years. Most existing approaches heavily rely on the explicit contributions (posts) of a user and overlook users' *implicit interests*, i.e., those potential user interests that the user did not explicitly mention but might have interest in. Given a set of active topics present in a social network in a specified time interval, our goal is to build an interest profile for a user over these topics by considering both explicit and implicit interests of the user. The reason for this is that the interests of free-riders and cold start users who constitute a large majority of social network users, cannot be directly identified from their explicit contributions to the social network. Specifically, to infer users' implicit interests, we propose a graph-based link prediction schema that operates over a representation model consisting of three types of information: user explicit contributions to topics, relationships between users, and the relatedness between topics. Through extensive experiments on different variants of our representation model and considering both homogeneous and heterogeneous link prediction, we investigate how topic relatedness and users' homophily relation impact the quality of inferring users' implicit interests. Comparison with state-of-the-art baselines on a real-world Twitter dataset demonstrates the effectiveness of our model in inferring users' interests in terms of perplexity and in the context of retweet prediction application. Moreover, we further show that the impact of our work is especially meaningful when considered in case of free-riders and cold start users.

1. Introduction

With the emergence and growing popularity of online social networks such as Twitter, many users extensively use social posts to express their feelings and views about a wide variety of social events/topics as they happen in real time. This has made social networks as a viable source of information about users' interests with regards to the current active topics/events (Abel, Gao, Houben, & Tao, 2011). For instance, when looking at Twitter data during November 2010, the rivalry between the two English Premier League football teams, Spurs and Arsenal is a topic that has attracted a lot of discussion and interest. The development of techniques that can automatically detect such topics and model users' interests towards them from user activities in social networks has become an emerging research area in the recent years, which has the potential to improve the quality of applications that work on a user modeling basis, such as filtering twitter streams (Kapanipathi, Orlandi, Sheth, & Passant, 2011), news recommendation (Abel et al., 2011; Meguebli, Kacimi, Doan, & Popineau, 2017), retweet prediction (Feng & Wang, 2013) and hashtag recommendation (Li, Jiang, Liu, Qiu, & Sun, 2017), among others.

Most existing approaches for detecting users' interests rely heavily on the explicit contributions (posts) of a user (Abel et al., 2011;

[*] Corresponding author.
E-mail address: kahani@um.ac.ir (M. Kahani).

https://doi.org/10.1016/j.ipm.2017.12.003
Received 29 September 2017; Received in revised form 8 December 2017; Accepted 11 December 2017
Available online 21 December 2017

Yang, Sun, Zhang, & Mei, 2012). In other words, to detect a user's interests, these approaches predominantly consider the content that the user has posted, shared, viewed or favorited on her social profile. However, they struggle to identify a user's interests if the user has not explicitly mentioned them. For example, consider the following tweets posted by a user, who we call 'Mary':

- "*The opportunity to go top of the Premier League will give Arsenal an extra incentive to beat Spurs, according to Wenger* http://bit.ly/chgPjO"
- "*Arsenal won't win with Wenger's policy. Spurs continue to exceed expectations*"

Based on the keywords explicitly mentioned in Mary's tweets, one could easily infer that she is interested in the rivalry between Spurs and Arsenal. The interests that are directly observable in a user's tweets are referred to as *explicit interests*. Expanding on this example, another topic that emerged in 2010 was Prince William's engagement. Looking at Mary's posts she never referred to this topic in her tweet stream. However, it is possible that Mary is British and is interested in both football and the British Royal family; although, she never explicitly tweeted about the latter. If that is in fact the case, then Mary's interest profile would need to include an interest pertaining to the British Royal family. We refer to these concealed user interests as *implicit interests*, i.e., topics that the user did not explicitly engage with but might have interest in. The identification of implicit interests has received less attention in the literature but is of significant importance for the purpose of accurate user modeling especially for users who are not active or only tend to be passive consumers of content and consequently their available textual content is sparse and does not reveal sufficient clues about their interests (Spasojevic, Yan, Rao, & Bhattacharyya, 2014).

The main objective of our work in this paper is to build a user interest profile by considering both explicit and implicit interests of the user in a given time interval from Twitter. Based on the homophily principle (McPherson, Smith-Lovin, & Cook, 2001), users tend to interact with users with common interests or preferences. Therefore, interactions between users can be considered to be an important clue for inferring their interests (Wen & Lin, 2010). In addition, relatedness between topics is another important aspect that can uncover implicit interests of users (Bhattacharya, Zafar, Ganguly, Ghosh, & Gummadi, 2014; Shen, Wang, Luo, & Wang, 2013). In this work, we combine these two factors into a unified heterogeneous representation model to consider them simultaneously.

More succinctly, the key contributions of our work are as follows:

- We model users' interests based on their inclination towards the active topics on Twitter. The literature is abundant with techniques that automatically detect active topics from social networks (Alvarez-Melis & Saveski, 2016). In our work, we develop techniques that determine, whether or not, and to what extent, a user is interested in these active topics on Twitter by considering both explicit and implicit interests of the user.
- We propose a graph-based link prediction framework to determine the so-called *implicit interests* of a given user. Our work considers a heterogeneous graph that includes three types of information: *i)* user explicit contributions to active topics, *ii)* the relatedness between those active topics, and *iii)* relationship between users to incorporate theory of homophily. To the best of our knowledge, the proposed framework is among the first to provide such a holistic approach for identifying users implicit interests.
- Another important contribution of our work is the investigation of whether it is possible to identify user interests for those users who are not highly active on the social network either because they are cold start users or free riders. This is especially important when considering the fact that studies have reported that most users on social networks can exhibit cold start or free riding behavior (Romero, Galuba, Asur, & Huberman, 2011). As such, methods that rely on explicit user information for determining user interests might not work well under these circumstances. In our experiments, we will additionally show that the impact of our work is especially meaningful when considered in such context.

The rest of the paper is organized as follows: Section 2 reviews the related work. The proposed approach is introduced in Sections 3. Section 4 is dedicated to the experiments and evaluation of the proposed model, which is followed in Section 5 by a discussion about our motivations and research findings. Finally, Section 6 sheds light on future work and concludes the paper.

2. Related work

Our work in this paper first extracts active topics from Twitter and then determines a given user's inclination towards these active topics. We assume that an existing state of the art technique such as those proposed in Alvarez-Melis and Saveski (2016) and Huang et al. (2017) can be employed for extracting and modeling active topics. Therefore, we will not be engaged with proposing a new method for the identification of topics and will only focus on determining the interest of users towards the topics once they are identified. Given this focus, we review the work related to the problem of user interest detection from social networks. Interested readers are encouraged to see Aiello et al. (2013), Farzindar and Khreich (2015) and Srijith, Hepple, Bontcheva, and Preotiuc-Pietro (2017) for the state of the art on topic and event detection.

Current approaches to user interest identification from social networks can be viewed as either single-source or multi-source (Abel, Herder, Houben, Henze, & Krause, 2013). In single-source approach, only one social network is considered as the source of information. Most of these works use Twitter as their source of information, because the information that the users publish on Twitter are more publicly accessible compared to other social networks. Multi-source approach, on the other hand, is based on the idea that a user has different profiles in different social networks, and to extract her interests more accurately, it would be better to extract and integrate her information from all those profiles (Spasojevic et al., 2014). Independent from how the information is collected and integrated, existing user interest detection methods can be broadly classified into two categories: *explicit* interest detection and *implicit* interest detection methods.

2.1. Explicit interest detection

There is a rich line of research on user interest detection from social networks that have focused on extracting *explicit* interests through analysing textual contents of users. For example, Yang et al. (2012) have modeled user interests by representing her tweets as a bag of words, and by applying cosine similarity to determine the similarity between the users in order to infer common interests. Weng, Lim, Jiang, and He (2010) have discovered user topics of interest by running Latent Dirichlet Allocation (LDA), the *de facto* standard in topic modeling, over the collection of a user's tweets. Xu, Lu, Xiang, and Yang (2011) have proposed a modified author-topic model where the latent variables are used to indicate whether the tweet is related to the user's (author) interests or not.

Since users in social networks can freely publish posts without any restriction, their posts are usually unstructured and include a nearly unlimited set of terms. By using *Bag of Words* and *Topic Modeling* techniques, user interest techniques might forgo the underlying semantics of the phrases in favor of highlighting the role of syntactical repetition of textual content (Kapanipathi, Jain, Venkatramani, & Sheth, 2014). Furthermore, these approaches assume that a single document contains rich information, as a result they may not perform so well on short, noisy and informal texts like tweets (Cheng, Yan, Lan, & Guo, 2014; Li, Wang, Zhang, Sun, & Ma, 2016). To address these issues, some recent work in interest identification have tried to utilize external knowledge bases to enrich the representation of short textual content and model user interests through semantic concepts linked to external knowledge bases such as DBpedia, YAGO and Freebase (Calegari & Pasi, 2013). For example, Abel et al. (2011) have proposed to enrich Twitter posts by linking them to related news articles and then extracting the semantic concepts mentioned in the enriched posts using web services provided by OpenCalais[1]. The identified semantic concepts are then used to build user interest profiles. Similarly, Kapanipathi et al. (2011) have modeled users' interests by annotating their tweets with DBpedia concepts, and have used these annotations for the purpose of filtering tweets. Michelson and Macskassy (2010) have proposed to extract a user's interests by first extracting a set of DBpedia concepts from her tweets and then identifying high-level user interests by traversing and analyzing Wikipedia category hierarchy for the extracted concepts. Similarly, Kapanipathi et al. (2014) have first extracted weighted primitive interests of a user as a bag of concepts extracted from the entities mentioned in the user's tweets. Then, the high-level interests of the user are extracted by mapping those primitive interests to the Wikipedia category hierarchy using a spreading activation algorithm.

Most of the works that use semantic concepts for representing user interests fall short when user interests do not necessarily have an exact corresponding semantic concept in the knowledge base. In other words, these models are successful to the extent they find a suitable concept to represent a user interest but they will not be able to model and detect users' interests if such interests are not formally represented in the knowledge base. For instance, on 16 November 2010, when Prince William's engagement was first mentioned on Twitter, it was received by a large number of users posting about it; however, at the time, no single DBpedia concept or Wikipedia article was available to link this topic to. Following our earlier work Zarrinkalam, Fani, Bagheri, Kahani, and Du (2015), for extracting explicit user interests, we view each topic of interest as a collection of several semantic concepts which are temporally correlated on Twitter, and model user interests over these topics. Therefore, even if a single corresponding semantic concept is not available in the external knowledge base, we construct its semantics by using existing concepts.

2.2. Implicit interest detection

While most of the existing works have focused on extracting explicit interests, there are some works that have been dedicated to inferring implicit interests of the users. For instance, Wang, Liu, He, and Du (2013) have used homophily to infer interests of a user based on the information provided by her neighbors. Based on the homophily principle, users tend to connect to users with common interests or preferences. Wang, Zhao, He, and Li (2014) have extended this principle by extracting user interests based on implicit links between users in addition to explicit relations. For example, if two users share many followers, they are likely to be similar in terms of their topical interests. Bhattacharya et al. (2014) have inferred topical expertise of famous Twitter users via their Twitter lists features and then discovered the interests of a user based on the topical expertise of the users that she follows. Their approach is based on the observation that a user generally follows the influential users related to her topical interest. He, Liu, He, Tang, and Du (2015) have also followed this observation and proposed a modified topic model to extract interest tags for non-famous Twitter users, based on their relationship with famous users.

While the above works incorporate the social relationship between users, they do not consider the relatedness between the topics of interest. In our previous work, to infer implicit interests of users, we have proposed a graph-based link prediction scheme to combine both the relatedness of the topics and social relationship between users into a unified representation model and illustrated its functionality as a proof of concept (Zarrinkalam, Fani, Bagheri, & Kahani, 2016). This current paper is substantially different from our previous work in the following aspects:

1. We take into account the heterogeneous nature of the representation model by applying link prediction methods customized for heterogeneous graphs in the context of implicit interest detection, which is novel in this paper. Our experiments validate the

[1] http://www.opencalais.com.

necessity to consider heterogeneity of our representation model to predict users' implicit interests.

2. In addition, we propose and incorporate relationship strengths within the representation model by using weighted edges in the graph and considering weighted link prediction, which is a new addition to this paper. Our findings highlight the fact that incorporating the strength of relationships in the representation model can contribute to the improvement of the performance of our implicit interest detection method.
3. Comprehensive experiments are conducted and new findings are reported in this paper. Specifically, we evaluate the effectiveness of the extracted user interest profiles in comparison with the state of the art in terms of perplexity and in the context of retweet prediction application. We further analyze the different user interest detection strategies by investigating the impact of different level of user engagement on Twitter, on their performance.

3. Proposed approach

The overarching objective of our work is to identify a user's interests towards the active topics on Twitter. A topic z has traditionally been defined as a semantically coherent theme which has received substantial attention from the users.

Let $\mathbb{Z} = \{z_1, z_2, \dots, z_K\}$ be K active topics, for each user $u \in \mathbb{U}$, we define her interest profile, which is the distribution of u's interests over $\mathbb{Z}$, as follows:

Definition 1 (*Interest Profile***)**. Given a set of topics $\mathbb{Z}$ and a set of users $\mathbb{U}$, an interest profile of user $u \in \mathbb{U}$, called $P(u)$, is represented by a vector of weights over K topics, i.e., $(f_u(z_1), \dots, f_u(z_K))$, where $f_u(z_k)$ denotes the degree of u's interest in topic $z_k \in \mathbb{Z}$. A user interest profile is normalized so that the sum of all weights in a profile is equal to 1.

The main problem addressed in our work is to build an interest profile for a user, $P(u) = (f_u(z_1), \dots, f_u(z_K))$, where $f_u(z_k)$ is calculated based on explicit and implicit interests of the user. We divide this problem into two subproblems: *explicit interest detection* and *implicit interest detection* in which the output of the first subproblem becomes the input of the second one. In this section, we concretely formulate these subproblems and propose our approach.

3.1. Detecting explicit user interests

The interests that are directly observable in a user's tweets are referred to as explicit interests. User explicit interest detection over active topics on Twitter have already been studied in the literature and therefore are not the focus of our work. We are able to work with any topic and interest detection method to extract topics $\mathbb{Z}$ and the explicit interest profile for each user u toward these topics, denoted as $P_E(u)$.

Latent Dirichlet Allocation (LDA) is one of the well-known unsupervised techniques used for identifying latent topics from a corpus of documents. However because it is designed for regular documents, it may not perform so well on short, noisy and informal texts like tweets and might suffer from the sparsity problem (Cheng et al., 2014; Derczynski et al., 2015). As proposed in Zarrinkalam et al. (2015), to obtain better topics from twitter without modifying the standard topic modeling methods, we enrich each Twitter micropost (tweet) m from our corpus $\mathbb{M}$ by using a semantic annotator and employ the extracted concepts, which can lead to the reduction of noisy content within the topic detection process. Therefore, in our work, each tweet is considered to be a set of one or more semantic concepts that collectively denote the underlying semantics of that tweet. As explained later in the experiments section, we employ TagMe (Ferragina & Scaiella, 2012) to link tweets to Wikipedia concepts. For instance, for a tweet such as *"Arsenal won't win with Wenger's policy. Spurs continue to exceed expectations"*, by using TagMe (Ferragina & Scaiella, 2012), we model it as a collection of three semantic concepts, namely "Arsenal_F.C",[2] "Arsene_Wenger" and "Tottenham_Hotspur_F.C". Therefore, we view a topic, defined in Definition 2, as a distribution over Wikipedia concepts (Zarrinkalam et al., 2015).

Definition 2 (*Topic***)**. Let $\mathbb{M}$ be a corpus of Twitter microposts and $\mathbb{C}$ be the vocabulary of Wikipedia concepts, a topic z, is defined to be a vector of weights, i.e., $(w_z(c_1), \dots, w_z(c_{|\mathbb{C}|}))$, where $w_z(c_i)$ shows the participation score of $c_i \in \mathbb{C}$ in forming topic z. Collectively, $\mathbb{Z} = \{z_1, z_2, \dots, z_K\}$ denotes a set of K topics extracted from $\mathbb{M}$.

To extract the topics from tweets by using LDA, documents should naturally correspond to tweets. However, since our goal is to understand the topics that each user u is interested in rather than the topics that each single tweet is about, similar to previous works in the literature (Weng et al., 2010), we aggregate the published or retweeted tweets of a user u in $\mathbb{M}$, i.e., $\mathbb{M}_u$, into a single document. LDA has two parameters to be inferred from the corpus of documents: topic-term distributions and document-topic distributions. Given that each document corresponds to a user u and Wikipedia concepts $\mathbb{C}$ as the vocabulary of terms, by applying LDA over the corpus of tweets $\mathbb{M}$, the results produce the following two artifacts:

- K topic-concept distributions, where each topic concept distribution associated with a topic $z \in \mathbb{Z}$ represents an active topic in $\mathbb{M}$, i.e., $(w_z(c_1), \dots, w_z(c_{|\mathbb{C}|}))$
- $|\mathbb{U}|$ user-topic distributions, where each user-topic distribution associated with a user u represents the explicit interest profile of user u, i.e., $P_E(u) = (f_u^E(z_1), \dots, f_u^E(z_K))$.

[2] http://en.wikipedia.org/wiki/Arsenal_F.C.

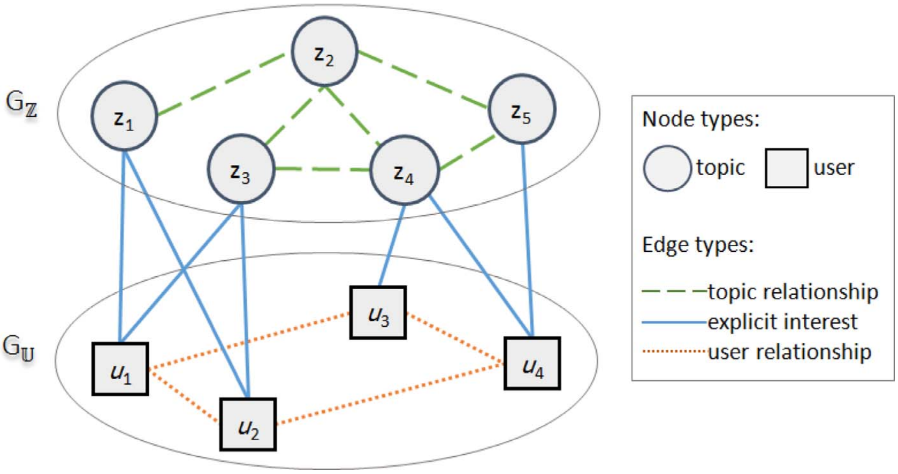

Fig. 1. An illustration of the representation model.

Building a user interest profile based only on explicit contributions of a user would not take the additional information about the social relationships and topics similarities into account and consequently could not be as effective for identifying user interests that have not been explicitly mentioned by the user, but are at the same time important for accurately modeling the user. In the next section, we describe our approach to benefit from such additional information for implicit interest detection.

3.2. Detecting implicit user interests

Having identified the explicit interest profile of each user u, i.e., $P_E(u) = (f_u^E(z_1), \ldots, f_u^E(z_K))$, our goal is to model and identify implicit interests of a user u in each topic z_k that she has not explicitly expressed interest in. To address this challenge, we propose to turn the problem of detecting implicit interests into a link prediction problem that operates over a heterogeneous graph including three types of information: *i)* users' explicit contributions to active topics, *ii)* the homophily principle, which states that users that are connected on the social network have a high likelihood of sharing similar interests, and *iii)* the possible relationship or similarity between the active topics themselves. In this section, we are interested in proposing a holistic approach for identifying users' implicit interests.

3.2.1. Representation model formalization

Our underlying representation model can be formalized as follows:

Definition 3 (*Representation Model***).** Given a set of topics $\mathbb{Z}$ and a set of users $\mathbb{U}$, our representation model, denoted by $G = (G_{\mathbb{U}} \cup G_{\mathbb{U}\mathbb{Z}} \cup G_{\mathbb{Z}})$, is a heterogenous graph composed of three subgraphs, $G_{\mathbb{U}}$, $G_{\mathbb{U}\mathbb{Z}}$ and $G_{\mathbb{Z}}$. $G_{\mathbb{U}}$ is based on homophily principle and represents the social relations between users on Twitter, $G_{\mathbb{U}\mathbb{Z}}$ represents explicitly observable user-topic relations, i.e., the explicit interests, and $G_{\mathbb{Z}}$ denotes the potential relationships between the active topics in $\mathbb{Z}$. An illustration of the representation model is shown in Fig. 1.

Definition 4 (*User Graph***).** The user graph is a weighted undirected graph $G_{\mathbb{U}} = (\mathbb{V}, \mathbb{E}, g)$ where $\mathbb{V} = \mathbb{U}$, and edges $\mathbb{E}$ are formed by observing a social relation between two users. The function $g(e_{u_i,u_j})$ shows the weight of the relation between two users u_i and u_j.

Welch, Schonfeld, He, and Cho (2011) have compared the retweeting relation to the followership relation and have concluded that retweeting is a much stronger indicator of topical interest. It has also reported in Wang et al. (2014) that if two users retweet the tweets of common users, they are likely to be similar in terms of their topical interests. As a result, we build the user graph based on the co-retweet similarity between the users. Formally, let $RT(u_i, u_k)$ be the number of tweets retweeted by user u_i from user u_k, we compute $g(e_{u_i,u_j})$ as follows (Wang et al., 2014):

$$g(e_{u_i,u_j}) = \frac{\sum_{u_k \in \mathbb{U}} RT(u_i, u_k) RT(u_j, u_k)}{\sqrt{\sum_{u_k \in \mathbb{U}} RT^2(u_i, u_k)} \sqrt{\sum_{u_k \in \mathbb{U}} RT^2(u_j, u_k)}} \tag{1}$$

Now, we introduce the formalization of the user-topic graph in our representation model as follows:

Definition 5 (*User-Topic Graph***).** The user-topic graph is a weighted undirected graph $G_{\mathbb{U}\mathbb{Z}} = (\mathbb{V}, \mathbb{E}, g)$ where $\mathbb{V} = \mathbb{Z} \cup \mathbb{U}$, and edges $\mathbb{E}$ are established by observing a user's explicit contributions towards active topic $\mathbb{Z}$. The weight function $g(e_{u,z})$ represents the degree of u's explicit interest in topic $z \in \mathbb{Z}$, i.e., $g(e_{u,z}) = f_u^E(z)$ which is calculated as described in Section 3.1.

The third type of information that we consider in our model is the relationship between topics. In other words, we are interested in knowing whether relationships between topics can be used to infer implicit interests of the users. For instance, one could potentially infer that a user might be implicitly interested in other topics that are similar to the topics that the user has explicit interest in. The topic subgraph in our representation model is built on this basis.

Definition 6 (*Topic Graph*). The topic graph is a weighted undirected graph $G_{\mathbb{Z}} = (\mathbb{V}, \mathbb{E}, g)$ where $\mathbb{V} = \mathbb{Z}$, and the edges in $\mathbb{E}$ represent the relationships between these topics. The weight function $g(e_{z_i,z_j})$ represents the degree of relatedness or similarity of the topics which is calculated using measures described in the following.

Topic Relatedness: Some researchers have already argued that users often have coherent and related interests (Bhattacharya et al., 2014; Shen et al., 2013). Based on this, we hypothesize that users are likely to be interested in topics that are conceptually similar to the topics that they have shown explicit interest in. For instance, users who have shown to be interested in the English Premier League might also be interested in the Spanish La Liga. Therefore, we consider topic relatedness to identify users' implicit interests. In order to model topic relatedness in $G_{\mathbb{Z}}$, i.e., $g(e_{z_i,z_j})$, we consider three different topic relatedness measures in our experiments, namely: *i)* semantics relatedness (S), *ii)* collaborative relatedness (C), and *iii)* hybrid relatedness (CS).

In the *semantic relatedness* approach, the relatedness of two topics is determined based on their constituent semantic concepts. In other words, two topics are considered to be similar to the extent that the semantic concepts that make up those topics are similar. Different methods have already been proposed in the literature to calculate the similarity between two concepts, e.g., through link structure analysis on Wikipedia (Duong, Nguyen, & Nguyen, 2016; Jiang, Bai, Zhang, & Hu, 2017; Jiang, Zhang, Tang, & Nie, 2015), among others; however, exploring the impact of these different similarity methods is beyond the scope of this paper. Given each topic z as a distribution over semantic concepts, i.e., $z = (w_z(c_1), \ldots, w_z(c_{|\mathbb{C}|}))$, we simply calculate the semantic relatedness of two topics by calculating the cosine similarity between their respective concept weight distribution vectors as shown in Eq. (2).

$$S(z_i, z_j) = \frac{\sum_{c\in\mathbb{C}} w_{z_i}(c) w_{z_j}(c)}{\sqrt{\sum_{c\in\mathbb{C}} w_{z_i}^2(c)}\sqrt{\sum_{c\in\mathbb{C}} w_{z_j}^2(c)}} \tag{2}$$

In the *collaborative relatedness* approach, the relatedness of two topics is determined based on a collaborative filtering strategy where relatedness is measured based on users' overlapping contributions toward these topics. Given a user-topic graph $G_{\mathbb{U}\mathbb{Z}}$, we regard the problem of computing the collaborative relatedness of topics as an instance of a model-based collaborative filtering problem (Adomavicius & Tuzhilin, 2005). To this end, we model the user-topic graph information as a user-item rating matrix R of size $|\mathbb{U}| \times |\mathbb{Z}|$, in which an entry in R, denoted by r_{uz}, is used to represent the weight of the edge between user u and topic z in user-topic graph $G_{\mathbb{U}\mathbb{Z}}$, i.e., $f_u^E(z)$. By considering matrix R as the ground-truth item recommendation scores, our problem is to learn the relationship between topics in the form of an item similarity matrix. We adopt a factored item-item collaborative filtering method (Kabbur, Ning, & Karypis, 2013) that learns item-item similarities (topic relatedness) as a product of two rank matrices, P and Q, which denote latent factors of items. In our model, the rating for a given user u on topic z_i is estimated as:

$$\hat{r}_{uz_i} = b_u + b_i + (n_u^+)^{-\alpha} \sum_{z_j\in R_u^+} p_j q_i^T \tag{3}$$

where R_u^+ is the set of topics that user u is interested in, p_j and q_i are the learned topic latent factors for the topics z_j and z_i, n_u^+ is the number of topics that user u is interested in and α is a user specified parameter between 0 and 1. According to Kabbur et al. (2013), matrices P and Q can be learnt by minimizing a regularized optimization problem:

$$minimize\left(\frac{1}{2}\sum_{u,z_i\in R}\left\| r_{uz_i} - \hat{r}_{uz_i}\right\|_F^2 + \frac{\beta}{2}(||P||_F^2 + ||Q||_F^2) + \frac{\lambda}{2}\left\| b_u \right\|_2^2 + \frac{\gamma}{2}||b_i||_2^2\right) \tag{4}$$

where the vectors b_u and b_i correspond to the vector of user u and topic z_i biases.

The optimization problem can be solved using Stochastic Gradient Descent to learn two matrices P and Q. Given P and Q as latent factors of topics, the collaborative relatedness of two topics z_i and z_j, i.e., $C(z_i, z_j)$, is computed as the dot product between the corresponding factors from P and Q, i.e., $C(z_i, z_j) = p_i.\, q_j^T$.

While the collaborative relatedness measure can find topic relatedness based on user's contributions to the topics, it overlooks the semantic relatedness between the two topics. Therefore, in the third approach, we develop a *hybrid* relatedness measure that considers both the semantic relatedness of the concepts within each topic as well as users' contributions towards the topics. We follow the assumption of Yu, Wang, and Gao (2014) for utilizing item attribute information to add the item relationship regularization term into Eq. (4). On this basis, two topic latent feature vectors would be considered similar if they are similar according to their attribute information. The topic relationship regularization term is defined as:

$$\frac{\sigma}{2}\sum_{i=1}^{|\mathbb{Z}|}\sum_{i'=1}^{|\mathbb{Z}|} S(z_i, z_{i'})(||q_i - q_{i'}\Big\|_F^2 + ||p_i - p_{i'}||_F^2) \tag{5}$$

where σ is a parameter to control the impact of topic concept information. In our proposed approach, attributes of each topic are its constituent concepts and $S(z_i, z_{i'})$ is calculated by measuring the semantic relatedness of two topics z_i and $z_{i'}$ based on Eq. (2).

After adding the item relationship regularization term (Eq. (5)) to Eq. (4), we learn P and Q as latent factors of topics, by minimizing the resulting regularized optimization problem. Similar to collaborative filtering relatedness method, the hybrid relatedness of two topics z_i and z_j, i.e., $CS(z_i, z_j)$, is computed as the dot product between the corresponding factors from P and Q, i.e.,

$CS(z_i, z_j) = p_i . q_j^T$.

We use the above three sub-graphs to build the representation model defined in Definition 3. Our model for inferring implicit interests of users operates over this representation model.

3.2.2. Implicit interest prediction

After building the representation model G, to build *Implicit Interest Profile* of user $u \in \mathbb{U}$, denoted as $P_I(u) = (f_u^I(z_1),...,f_u^I(z_K))$, we infer her implicit interest toward each topic $z_k \in \mathbb{Z}$ that she has not explicitly expressed interest in, by formulating a graph-based link prediction problem that operates over G. We follow two possible solutions for handling the link prediction problem: *i) Homogeneous Link Prediction*: treating all types of nodes and edges equally and applying link prediction strategies designed for homogeneous networks, *ii) Heterogeneous Link Prediction*: considering strategies which are customized for heterogeneous networks. In the following we describe these two solutions.

It is important to note that because the edge weights in different subgraphs, i.e., $G_{\mathbb{U}}$, $G_{\mathbb{Z}}$ and $G_{\mathbb{UZ}}$ come from different sources, they have different scales. Therefore, in the experiments, before applying the link prediction strategies, we scale the weights of each subgraph to the range [0, 1], separately.

Homogeneous Link Prediction: Most of the unsupervised link prediction strategies that operate on homogeneous graphs generate $score_{xy}$ for a pair of nodes (x, y) based on the neighborhood or path information of these nodes (Liben-Nowell & Kleinberg, 2007). Neighborhood methods are based on the idea that two nodes x and y are more likely to have a link if they have many common neighbors. Path-based methods consider the ensemble of all paths between two vertices. Both methods are based on a predictive score function for ranking links that are likely to occur. According to the experiments done in Liben-Nowell and Kleinberg (2007), there is no single superior method among existing work and their quality is dependent on the structure of the underlying graph. Therefore, in our experiments, we exploit some of the well-known and frequently used link prediction strategies that can be applied in weighted graphs for inferring implicit interests of a user.

From the neighborhood-based methods, we adopt Common Neighbors and Adamic/Adar (Adamic & Adar, 2003) measures. Common neighbors is defined as the number of common neighbors shared by two nodes x and y while Adamic/Adar computes the similarity between two nodes x and y by looking at their common neighbors' features and weighting rarer features more heavily. From the path-based methods, we employ Katz (1953) and PropFlow (Lichtenwalter, Lussier, & Chawla, 2010) measures. Katz is a weighted summation of counts of paths between two nodes, exponentially damped by length to count short paths more heavily and PropFlow is a random walk-based measure that assigns the weights to each path using the products of proportions of the flows on the edges. Table 1 introduces the measures to calculate the weighted (Lu & Zhou, 2009; de Sa & Prudêncio, 2011) and unweighted (Liben-Nowell & Kleinberg, 2007) versions of each link prediction strategy.

Given the representation model G and in order to infer implicit interest of a user u toward topic $z \in \mathbb{Z}$, i.e. $f_u^I(z)$, we adopt one of the unsupervised link prediction strategies introduced in Table 1 to compute a connection weight $score_{u,\ z}$ for the pair (u, z) in G. We then assign $score_{u,\ z}$ to $f_u^I(z)$.

Heterogeneous Link Prediction: As previously mentioned, most of the existing link prediction methods, e.g., the methods listed in Table 1, are defined for homogeneous graphs and treat all types of nodes and edges equally. However, as our representation model is a heterogeneous graph, the neighbors of a node could belong to multiple types and the paths between two nodes could have different meanings. Sun, Han, Yan, Yu, and Wu (2011b) proposed the concept of *meta-path* for heterogeneous information network analysis, which is now widely known and used in different data mining tasks (Shi, Li, Zhang, Sun, & Yu, 2017) such as ranking (Liu, Yu, Guo, & Sun, 2014a), clustering (Sun, Aggarwal, & Han, 2012) and link prediction (Cao, Kong, & Yu, 2014).

Table 1

Values for $score_{xy}$ under four link prediction strategies chosen for user implicit interest prediction. $\Gamma(x)$ denotes the set of neighbors of vertex x and w_{xy} denotes the weight of the edge between nodes x and y.

	Unweighted	Weighted
Common Neighbors	$\Gamma(x) \cap \Gamma(y)$	$\sum_{k \in \Gamma(x) \cap \Gamma(y)} w_{xk} + w_{yk}$
Adamic/Adar	$\sum_{k \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log \mid \Gamma(k) \mid}$	$\sum_{k \in \Gamma(x) \cap \Gamma(y)} \frac{w_{xk} + w_{yk}}{log\left(1 + \sum_{a \in \Gamma(k)} w_{ka}\right)}$
Katz	$\sum_{\ell=1}^{\infty} \beta^{\ell} \lvert path_{x,y}^{<\ell>} \rvert$	$\sum_{\ell=1}^{\infty} \beta^{\ell} \sum_{p \in path_{x,y}^{<\ell>}} \frac{\sum_{(x,y) \in p} w_{xy}}{\ell}$
	$path_{x,y}^{<\ell>}$: a set of all paths with length ℓ from x to y β: damping factor to give the shorter paths more weights	
PropFlow	If nodes x and y are directly linked: $score_{x,y} = score_{a,x} \frac{1}{\lvert H(x) \rvert}$	$score_{x,y} = score_{a,x} \frac{w_{xy}}{\sum_{k \in H(x)} w_{xk}}$
	where a is the previous node of x on a random walk path, and $H(x)$ is a set of x's neighbors whose depth is greater than the depth of x from the starting node. If x is the starting node, $score_{a,x} = 1$.	
	If x and y are indirectly linked, $score_{x,\ y}$ is the sum of PropFlow through all the shortest paths from x to y.	

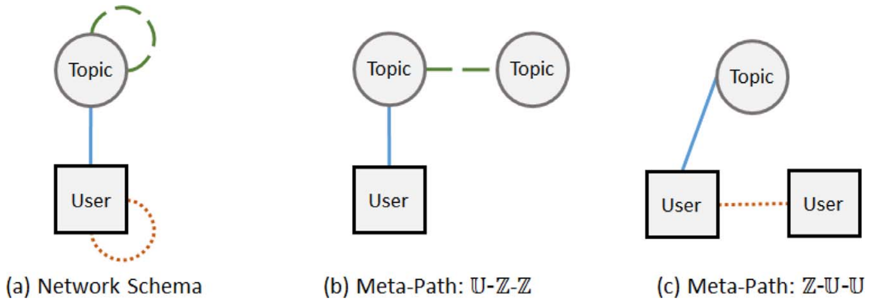

Fig. 2. network schema of the representation model and two sample meta-paths.

For instance, in order to solve the problem of link prediction in heterogeneous graphs, Sun, Barber, Gupta, Aggarwal, and Han (2011a) have proposed a new methodology called *PathPredict*, i.e., the meta-path-based relationship prediction model. A meta-path is a path defined over the heterogeneous network schema which can be used to define topological features with different semantic meanings. Therefore, to distinguish different types of nodes and edges, following the work in Sun et al. (2011a), we use the PathPredict methodology to infer implicit interest of a user u toward each topic $z \in \mathbb{Z}$. Fig. 2(a) summarizes our representation model using a meta structure called network schema. As shown in this figure, our representation model contains two types of nodes: user ($\mathbb{U}$) and topic ($\mathbb{Z}$).

Based on the PathPredict approach, for the target relation $< \mathbb{U}, \mathbb{Z} >$, any meta-paths starting with type $\mathbb{U}$ and ending with type $\mathbb{Z}$ other than the target relation itself can be used as the topological features. We extract all of these meta-paths by traversing on our network schema using Breadth First Search (BFS) within a fixed length constraint. Generally speaking, the number of possible meta-paths for a given heterogeneous graph grows exponentially with their maximum path length l_{max}. As pointed in Sun et al. (2011b), meta-paths with relatively short length are good enough for capturing the structure of heterogeneous graphs, and a long meta-path may even reduce quality. On this basis, we extract short meta-paths with a maximum path length of $l_{max} = 3$.

For example, two automatically determined meta-paths of length 2 retrieved from our network schema are illustrated in Fig. 2 (b) and (c). The meta-path $\mathbb{U} - \mathbb{Z} - \mathbb{Z}$ (i.e., user-topic-topic) considers the related topics to explicit interests of a user as her implicit interests and the meta-path $\mathbb{U} - \mathbb{U} - \mathbb{Z}$ (i.e. user-user-topic) infers the implicit interests of a user based on the homophily principle, which states that users that are connected on the social network have a high likelihood of sharing similar interests.

Once the meta-paths are retrieved from the network schema, for each user-topic pair (u, z) in the representation model G, we can use Path Count based or Random walk based measures (Sun et al., 2011a) to quantify each meta-path as topological features. In this paper, without loss of generality, we use *Path Count (PC)* as the default measure and the weighted version of this measure called *Weighted Path Count (WPC)*. For a given meta-path, Path Count simply counts the number of meta-path instances in the representation model starting at u and ending at z. Similarly, Weighted Path Count considers the edge weights by summing the weights of edges included in each meta-path instance.

Given the training user-topic pairs and the extracted topological features for them, a logistic regression model is trained as the learning framework to learn the weights associated with these features. Then, for a given test pair (u, z), we apply the learned coefficients to the topological features to predict the implicit interest of user u toward topic $z \in \mathbb{Z}$, i.e., $f_u^I(z)$.

4. Experiments

In this section, we describe our experiments in terms of the dataset, setup and performance compared to the state of the art.

4.1. Dataset and experimental setup

We use the publicly available Twitter dataset[3]Abel et al. (2011) that includes about 3M tweets posted by approximately 130,168 users, starting from Nov. 1st and lasting for two months until Dec. 31st 2010. Fig. 3 depicts the number of tweets per user in our dataset. Twitter datasets suffer from participation inequality, where a minority of users usually contribute the most while the others just free-ride. As shown in Fig. 3, only 15% of the users contribute more than 16 tweets within the two month period and the other users have less than 16 tweets. This reinforces the underlying purpose of our work, which is to identify users' interests when they are not expressed; for instance, in the case of cold start of free-rider users.

In order to annotate the text of each tweet, we use the TAGME RESTful API[4] with the recommended scoring threshold of 0.1 which resulted in 350,731 unique concepts. The choice of TagMe is motivated by a recent study that has shown this semantic annotator performs very reasonably on different types of text such as tweets, queries and web pages (Cornolti, Ferragina, & Ciaramita, 2013). The number of concepts per tweet is shown in Fig. 4. From the tweet content perspective, the complementary cumulative distribution of the concepts shows that in more than 85% of the tweets, the tweets included at least one semantic concept.

[3] http://wis.ewi.tudelft.nl/websci11.

[4] https://tagme.d4science.org/tagme/.

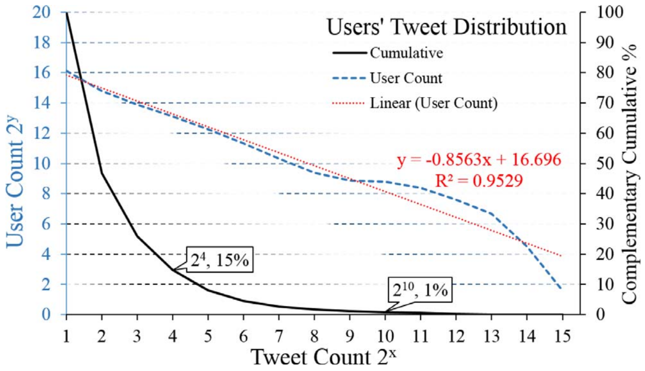

Fig. 3. The number of tweets per user and its complementary cumulative distribution.

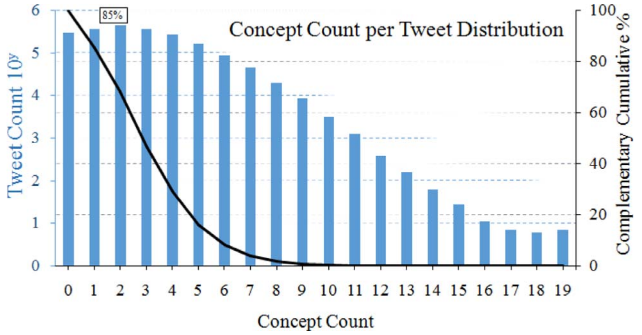

Fig. 4. The number of concepts per tweet and its complementary cumulative distribution.

Then, to extract topics $\mathbb{Z}$ of a given time interval T, and explicit interest profile of each user u over these topics, i.e., $P_E(u)$, as described in Section 3.1, we aggregate all concepts extracted from tweets of each user published in time interval T, into a single document and apply the Gensim implementation of LDA[5] with its default parameter settings on the collection of such documents. Similar to Wang et al. (2014), in our experiments, the length of the time interval T has been set to one month. Further, given LDA requires the number of topics K to be known a priori, we repeated all of our experiments on different number of topics: 50, 75 and 100. Because, The conclusions drawn from the results are similar in different number of topics in our experiments, we only report the experimental results obtained by setting $K = 50$.

As mentioned in Section 3.2.1, in order to compute the collaborative relatedness between topics, we learn the relationship between topics by adopting a factored item-item collaborative filtering method (Kabbur et al., 2013). In the learning step, we use the default parameter settings of the Librec library[6] and set $\beta = \gamma = \lambda = \sigma = 0.001$. Further the learning rate is set to 0.01, the number of item latent factors is set to 10 and we set the number of iterations to 1,000.

As mentioned in Section 3.2.2, we apply both homogeneous and heterogeneous link prediction strategies to infer implicit interests of users. In case of homogeneous strategies introduced in Table 1, we use the implementations made available on LPmade (Lichtenwalter & Chawla, 2011). Further, the implementation of PathPredict as a heterogeneous link prediction strategy has been provided by its authors (Sun et al., 2011a).

4.2. Analysis of the proposed implicit interest detection approach

In this section, comprehensive analyses are conducted to determine which factor or combination of factors are more influential in accurately predicting the implicit interests of users on Twitter. More specifically, we are interested in answering the following research questions to find the best configuration for our proposed user implicit interest detection approach among different possible design spaces:

RQ1. How and to what extent do different types of information present in our representation model facilitate the identification of users' implicit interests on Twitter?

RQ2. Does considering heterogeneity of the representation model affect prediction accuracy?

RQ3. Does considering weights of edges between users/topics in the representation model contribute to improved prediction accuracy?

[5] https://radimrehurek.com/gensim/models/ldamodel.html.

[6] http://www.librec.net/.

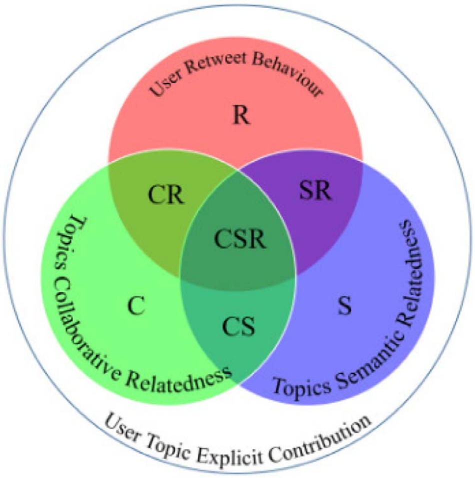
User Retweet Behaviour
R
CR
SR
CSR
Topics Collaborative Relatedness
C
CS
S
Topics Semantic Relatedness
User Topic Explicit Contribution

Fig. 5. Seven variants of the representation model.

4.2.1. *Evaluation methodology and metrics*

Our evaluation strategy is based on the 10-fold cross validation protocol. We first randomly divide all node pairs (*u, z*), such that $u \in \mathbb{U}$ and $z \in \mathbb{Z}$, into 10 equal sized subsamples and then pick one subsample for test and the rest of 9 subsamples for training. We repeat this procedure 10 times, with each of the subsamples used exactly once as the test data, thus producing 10 results. To evaluate the results, we compare them with the test set using the Area Under Receiver Operating Characteristic (AUROC) (Bradley, 1997). All the results in this section are the average results using 10-fold cross-validation.

4.2.2. *Results and discussion*

To answer the first research question, we define and analyze different variants of our representation model by varying its two main variation points: *i)* user relationships based on their retweeting behaviour (*R*) and *ii)* the type of topic relatedness measure used to compute topic relationships, i.e., semantic (*S*), collaborative (*C*) or hybrid (*CS*). By selecting and combining the different alternatives, we obtain seven variants that we will systematically compare in this section. These seven variants are illustrated in Fig. 5. The user's explicit interest information is included in all of the seven variants. As a brief example on how to interpret the model acronyms, Model *R* only uses user relationships in addition to user explicit interest information. Model *SR* considers topic relationships computed using the semantic relatedness measure, in addition to user retweet information and user's explicit interests. The rest of the models can be interpreted similarly.

To answer the second question, different link prediction strategies, introduced in Section 3.2.2, for both homogeneous and heterogeneous graphs are applied on all variants of our representation model. Further, to investigate the impact of edge weights on the accuracy of prediction (i.e., to answer the third question), we repeat the experiments for both the weighted and unweighted version of each link prediction strategy. The results are reported in Table 2 in terms of AUROC.

As mentioned earlier, Model *R* only considers retweeting behaviour between users in addition to users' explicit interests to infer users' implicit interests. Instead, models *S, C* and *CS* employ three different techniques for identifying topic relationships: model *S* uses semantic relatedness of the topics, model *C* uses collaborative relatedness and, model *CS* follows a hybrid approach. As depicted in Table 2, by applying all the selected link prediction strategies, in most of the cases, all these three models outperform Model *R* in terms of AUROC. This means that considering the relationships between the topics is a better clue for inferring implicit interests in comparison to when only user relationships are used.

As another observation, by comparing the results of applying all the link prediction strategies over the variants *S, C* and *CS*, it can

Table 2
The AUROC values showing the performance of different model variants. U denotes **U**nweighted and W denotes **W**eighted version of each link prediction method.

Link Prediction Strategies			R	S	SR	C	CR	CS	CSR
Homo	Common Neighbor	U	0.547	0.685	0.673	0.572	0.587	0.591	0.594
		W	0.547	0.745	0.726	0.562	0.58	0.582	0.588
	Adamic/Adar	U	0.547	0.684	0.672	0.574	0.588	0.591	0.594
		W	0.551	0.743	0.723	0.564	0.582	0.583	0.589
	Katz $\ell = 5$, $\beta = 0.05$	U	0.659	0.678	0.675	0.673	0.652	0.681	0.69
		W	0.619	0.583	0.591	0.563	0.578	0.581	0.596
	PropFlow $\ell = 5$	U	0.654	0.685	0.654	0.676	0.629	0.677	0.678
		W	0.659	**0.75**	0.662	0.656	0.605	0.657	0.657
Hetero	PathPredict	U	0.723	0.766	0.772	0.735	0.74	0.746	0.753
		W	0.746	0.785	**0.802**	0.747	0.749	0.755	0.761

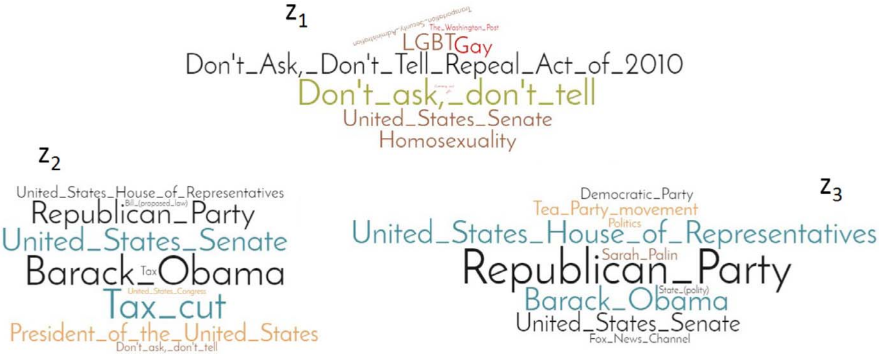

Fig. 6. Samples of related topics based on the semantic relatedness model.

be observed that using semantic relatedness for identifying topic relationships results in higher accuracy for the prediction of implicit interests, compared to the collaborative relatedness and hybrid measures. Therefore, semantic relatedness of topics is a more accurate indication of the tendency of users towards topics compared to collaborative relatedness of topics. Our explanation for this is that Twitter users are mostly focused on semantically coherent topics, i.e., they seem to follow topics that are from similar domains or genres. This is an observation that is also reported in Bhattacharya et al. (2014) and can be seen in the Who Likes What system.[7] For example, three samples of semantically related LDA topics extracted from our dataset are visualized in Fig. 6. Based on the top 10 concepts of each topic, the first topic z_1 is related to the act of repealing “don't ask, don't tell” policy in 2010. The second topic z_2 refers to the event of signing the tax cut law by Barak Obama in December 2010, and finally, the third topic z_3 is about the United States midterm elections in 2010. Since these topics are all about related important political events in the United States, it is easy to see that a user who is explicitly interested in one of these topics, might also be interested in the other two.

By comparing *C* and *CS*, it can be concluded that adding semantic relatedness for computing collaborative relatedness of topics leads to improved accuracy compared to using only collaborative relatedness alone. However, it is worth noting that considering only semantic relatedness outperforms the hybrid *CS* measure. The observation that *S* provides the best performance for predicting implicit interests is more appealing when the computational complexity involved in its computation is compared with the other methods. The computation of *S* only involves the calculation of the similarity of each pair of topics based on their constituent concepts, which is quite an inexpensive operation, whereas the computation of *C* and *CS* requires solving an optimization problem through Stochastic Gradient Descent, and hence they are computationally expensive tasks.

Based on the above observations, our observation with regards to RQ1 is that topic relatedness enables more accurate inference of implicit interests of users compared to social relations. Further, by comparing different types of topic relatedness measures, the results show that semantic relatedness measure is more effective compared to the collaborative filtering based method.

Furthermore, as explained in Section 3.2.2, given the proposed representation model, we adopted two approaches to predict implicit interests of a user. In the first approach, we simply consider the representation model as a homogeneous graph and utilize some of the well-known unsupervised link prediction strategies, i.e., Common Neighbor, Adamic/Adar, Katz and PropFlow. In the second one, we take into account the heterogeneity of our representation model and apply PathPredict link prediction strategy which is customized for heterogeneous graphs. Based on the results in Table 2, we can see that for all the variants of our representation model, the second approach improves the prediction accuracy of user's implicit interests compared to the first one in terms of AUROC.

Based on the results in Table 2, when we apply PathPredict, the variants of our representation model that incorporate both user relations and topic relations in addition to user explicit interests, i.e., *SR, CR* and *CSR*, outperform the variants that only use one kind of relation. For example, *SR* outperforms both *S* and *R*, similarly *CR* performs better than both *C* and *R*. However, this observation cannot be generalized to the results when we treat the representation model as a homogeneous graph. For instance, by comparing *S* and *SR*, when we apply homogeneous predictors, *SR* performs worse than *S*.

These findings relate back to RQ2 and validate the necessity to consider heterogeneity of our representation model to predict users implicit interests. Further, it highlights the fact that by utilizing heterogeneity of the representation model, we can better explore the effectiveness of incorporating both homophily theory and relationships between topics.

To investigate the impact of considering edge weights in our representation model, we compare the results of weighted and unweighted version of different link prediction methods over different variants of the representation model. As depicted in Table 2, it can be observed that for all of the representation models, by applying the weighted version of PathPredict, which is a supervised strategy, prediction accuracy is enhanced in terms of AUROC. However, by comparing weighted and unweighted version of the unsupervised link prediction methods, i.e., Common neighbor, Adamic/Adar, Katz and PropFlow, no generalizable observations can be made in any of the cases. For example, edge weights improve accuracy in Model *S*; however, accuracy is reduced when weights are considered in Model *C*. This observation is in line with the results reported in the literature that investigate the impact of edge weights on the accuracy of predictions in the unsupervised and supervised link prediction strategies (Lu & Zhou, 2009; de Sa &

[7] http://twitter-app.mpi-sws.org/who-likes-what/.

Prudêncio, 2011). Based on earlier studies, utilizing edge weights in unsupervised link predictors is a controversial issue and it depends on the underlying graph. In some case studies, using weights harmed the performance and in some cases it led to improvements. In contrast, it has been shown in de Sa and Prudêncio (2011) that the edge weights can improve the prediction results for *supervised* link prediction strategies.

Based on the results in Table 2 we can address RQ3 in that by applying the PathPredict link prediction strategy incorporating the strength of relationships in the representation model contributes to the improvement of the performance of our implicit interest detection method for all variants of our representation model.

In summary, as highlighted in Table 2, in the context of homogeneous link prediction strategies, model *S*, which relies solely on the semantic topic relatedness and user's explicit contributions shows the best performance when the weighted version of PropFlow is employed. Further, when we apply PathPredict as a heterogeneous link prediction strategy, we can see that *SR* which includes both topics and user relationships shows the best performance. Therefore, among different variants analysed in this section, two configurations, i.e., (*S*, Weighted PropFlow) and (*SR*, Weighted PathPredict), are selected for inferring implicit user interests and are used in the forthcoming experiments.

4.3. Analysis of user interest profile detection approaches

The main objective of our work in this paper is to extract a user interest profile by incorporating both explicit and implicit interests of the users from Twitter. In this section, to demonstrate the usefulness and effectiveness of our proposed work, we compare our proposed approach and other state-of-the-art baselines in terms of *Perplexity* as well as in the context of *retweet prediction application*. In each experiment, we also investigate the impact of the level of user engagement on Twitter (i.e., the number of tweets posted by a user) on the performance of the different models. We do this to be able to distinguish between cold start and free rider users and those users who have high engagement with the social networking platform.

4.3.1. Comparison methods

In the experiments, we consider the following user interest detection methods that can be applied over the identified active topics on Twitter for comparison:

EUI: In this method, the **E**xplicit **U**ser **I**nterest detection method described in Section 3.1 is used to build user interest profiles. This method estimates explicit user interests of each user u, i.e., $P_E(u) = (f_u^E(z_1), \ldots, f_u^E(z_K))$, based on the tweets that she has published or retweeted, without taking into account the social relations or topic relatedness.

TWang's Model (Wang et al., 2013): In this method, the interests of a user are inferred based on the posts published by the user, and her social connections. The proposed approach by Wang et al. (2013) is based on the idea that the interests are propagated to a user based on the influence she gets from her friends. The authors build a propagation graph based on different link information between users, and learn the user interest profile of user u_i, i.e., $P(u_i) = (f_{u_i}(z_1), \ldots, f_{u_i}(z_K))$ such that $f_{u_i}(z_k)$ is calculated as follows:

$$f_{u_i}(z_k) = \alpha \sum_{u_j \in \mathbb{U}} w(u_j, u_i) \times f_{u_j}(z_k) + f_{u_i}^E(z_k) \tag{6}$$

$$w(u_j, u_i) = \frac{RT(u_i, u_j)}{\sum_{u_k \in \mathbb{U}} RT(u_i, u_k)} \tag{7}$$

where α indicates the decay factor of influence from other users, and as reported in Wang et al. (2013) it is set to 0.5. Further, let $RT(u_i, u_j)$ be the number of tweets retweeted by user u_i from user u_j, the weight of the edge from node u_j to node u_i in the propagation graph, i.e. $w(u_j, u_i)$, is calculated based on Eq. (7). Finally, $f_u^E(z_k)$ denotes the explicit interests of user u in topic z_k and is calculated as described in Section 3.1.

JWang's Model (Wang et al., 2014): We also compare our model against one of the most related work in the literature by Wang et al. (2014), which proposes a regularization framework based on link structure assumption under which node similarities are evaluated based on the local link structures instead of only explicit links between nodes. The authors learn the user interest profile of user u_i, i.e., $P(u_i) = (f_{u_i}(z_1), \cdots, f_{u_i}(z_K))$ such that $f_{u_i}(z_k)$ is calculated as follows:

$$f_{u_i}(z_k) = \frac{\alpha \sum_{u_j} w_A(u_i, u_j) f_{u_j}(z_k) + \beta \sum_{u_j} w_H(u_i, u_j) f_{u_j}(z_k) + \lambda f_{u_i}^E(z_k)}{\alpha \sum_{u_j} w_A(u_i, u_j) + \beta \sum_{u_j} w_H(u_i, u_j) + \lambda} \tag{8}$$

where $w_A(u_i, u_j)$ is co-retweet similarity of two users u_i and u_j calculated based on Eq. (1), similarly $w_H(u_i, u_j)$ is co-retweeted similarity of u_i and u_j and $\alpha = \beta = 0.05$ and $\lambda = 0.9$ (Wang et al., 2014).

EIUI: This method is our approach that builds user interest profiles based on combining **E**xplicit and **I**mplicit **U**ser **I**nterest profiles. let $P_E(u) = (f_u^E(z_1), \ldots, f_u^E(z_K))$ be the explicit interests of a user u, in this method, we infer implicit interest of user u in each topic z_k that she is not explicitly interested in, i.e., the value of $f_u^E(z_k)$ is equal to 0. We build the implicit interest profile of each user u, i.e, $P_I(u) = (f_u^I(z_1), \ldots, f_u^I(z_K))$, using the method described in Section 3.2. Then, we combine the two profiles to build the interest profile of the user u, $P(u)$, simply by adding their corresponding vectors together. Based on the results reported in Section 4.2, two configurations are selected for user implicit interest prediction: (*S*, Weighted PropFlow) from the homogeneous variants, and (*SR*, Weighted PathPredict) from the heterogeneous variants. Consequently, we have two variants of our user interest detection approach in our experiments EIUI (*S*, Weighted PropFlow) and EIUI (*SR*, Weighted PathPredict).

Table 3
The performance of comparison methods in terms of perplexity. Lower value indicates better performance.

Method	K= 50	K= 75	K= 100
EUI	2220	2374	2529
TWang's Model	1716	2037	1918
JWang's Model	1551	1646	1594
EIUI (S, Weighted PropFlow)	1408	1410	1409
EIUI (SR, Weighted PathPredict)	**1348**	**1363**	**1358**

4.3.2. Evaluation by perplexity

Adopted from Wang et al. (2014), Xu et al. (2011), we utilize the perplexity metric to evaluate the overall generalizability of modeling unseen/implicit user interests in each comparison method. Perplexity is widely used in language modeling to evaluate the predictive power of a model (Steyvers, Smyth, Rosen-Zvi, & Griffiths, 2004). To do so, we consider the tweets published in November and December 2010 as the training and test data, respectively. As required for calculating perplexity, for each user u, we build her interest profile $P(u) = (f_u(z_1), \ldots, f_u(z_K))$ based on the training data and aggregate her tweets published in December 2010 as her test document indicating her unseen/implicit interests. Let D_{test} be the test set including the test documents of all the users, perplexity is computed as Wang et al. (2014):

$$Perplexity(D_{test}) = exp\left\{-\frac{\sum_{d_u \in D_{test}} \sum_{c \in \mathbb{C}_{d_u}} log\left(\sum_{z \in \mathbb{Z}} w_z(c) f_u(z)\right)}{\sum_{d_u \in D_{test}} |\mathbb{C}_{d_u}|}\right\} \tag{9}$$

where d_u is the test document of user u, $\mathbb{C}_{d_u}$ is the set of concepts in d_u and $w_z(c)$ is the probability of concept c in topic $z \in \mathbb{Z}$. A desirable method is one that achieves a smaller perplexity value on the test document set.

The perplexity results are presented in Table 3 by setting the number of topics to 50, 75 and 100. In contrast to the EUI method which is solely based on the textual content of user's tweets, all other comparison methods, i.e., EIUI, TWang and JWang's model utilize other information types available on Twitter such as user's social interactions and the relatedness between topics in order to extract users' interests. It can be observed that all these models outperform EUI by achieving lower perplexity. This observation indicates that incorporating other information types in addition to textual content of users leads to improved predictive power of the method for inferring user interests.

Based on the results in Table 3, it can be observed that the two variants of our proposed approach method EIUI for building user interest profiles which are based on user's implicit and explicit interests outperform both TWang's model (Wang et al., 2013) and JWang's model (Wang et al., 2014), which are two state of the art works in user interest detection from Twitter. It is important to note that both JWang's model and TWang's model are based on the social connections between users and do not take into account the relationship between topics. In other words, they overlook the fact that users might be predominantly interested in topics that are related to each other. This points to the importance of incorporating relatedness between topics to infer implicit/unseen user interests which is also confirmed in our experimental results in Section 4.2 by comparing different variants of our proposed model.

As another observation, it can be seen that JWang's model outperforms TWang's model in terms of Perplexity. The difference between them is that JWang's model also leverages implicit relationships between users in addition to their direct relations, which is not captured in the retweet graph. It shows that considering implicit links between users to infer users interests leads to more accurate results compared to influence propagation algorithms that are solely based on direct links. This observation is also reported in Wang et al. (2014). It is important to note that, following the results reported in Wang et al. (2014), our proposed framework, i.e., EIUI, also utilizes the benefits of using implicit links to infer user's interests by considering the co-retweet similarity between users.

Based on the results in Section 4.2, by comparing different variants of our proposed framework, we conclude that when we consider our representation model as a heterogenous graph and apply PathPredict link prediction method, which in effect distinguishes between different types of nodes and edges, the *SR* model yields the best performance in terms of the accuracy of predicting implicit interests. This observation is also confirmed in this section in terms of perplexity, by comparing two variants of our proposed framework. EIUI (*SR*, Weighted PathPredict) outperforms EIUI (*S*, Weighted PropFlow).

Cold Start User Problem: To investigate the impact of the user engagement level in posting tweets, on the performance of the interest detection methods, we partition the sample users into three groups based on the number of tweets they have published or retweeted in the training time interval, i.e., November 2010. The first group, referred to as *cold users*, consists of those users who have posted less than 30 tweets in November 2010 (i.e. less than one tweet per day on average). The second group (*semi-active users*) includes the users who have posted between 30 and 100 tweets. Finally, the users with more than 100 tweets in the time interval are referred to as *active users*.

Now, for each user group, we have calculated the performance of the comparison methods in terms of perplexity based on Eq. (9). The results are reported in Table 4 when the number of topics is set to $K = 50$.

Based on the results reported in Table 4, our proposed approach, i.e. EIUI, which is based on both the users' explicit and implicit interests outperforms all the baselines in terms of perplexity for all the user groups. We can conclude that regardless of the user's level of activity, our proposed approach reports better results compared to the baselines. It is worth noting that our proposed approach has a higher impact in case of the two first user groups where users have published less tweets compared to the active users. To make it clearer, the percentage of improvement of the two versions of our proposed approach, i.e., EIUI (*S*, Weighted PropFlow) and EIUI (*SR*,

Table 4
The performance of comparison methods for different user groups in terms of perplexity. The number of LDA topics K = 50.

Method	Cold users	Semi-active users	Active users
EUI	3141	2202	1985
TWang's Model	1849	1687	1751
JWang's Model	1552	1531	1653
EIUI (S, Weighted PropFlow)	1415	1371	1553
EIUI (SR, Weighted PathPredict)	**1352**	**1341**	**1493**

Weighted PathPredict), over the baselines for each user group are illustrated in Figs. 7 and 8, respectively. For instance, as depicted in Fig. 7, the EIUI (*S*, Weighted PropFlow) method improves the EUI method for cold users and semi-active users by a margin of 54.95% and 37.73%, respectively. However, our improvement in case of active users is 21.76% which is less compared to the cold users. This can be a sign that our proposed approach is more effective for cold users compared to active users.

4.3.3. Evaluation through retweet prediction

Inferring user interests from social networks plays an important role for many applications in the fields of information retrieval and recommender systems. Several researchers, such as Wang et al. (2014), Abel et al. (2011) and Zarrinkalam et al. (2015), have already suggested that the performance of user interest detection methods can be measured through observations made at the application level. In this section, adopted from Wang et al. (2014), we deploy a retweet prediction application and compare different user interest detection strategies by evaluating their impact on retweet prediction application.

There are different works for retweet prediction (tweet recommendation) that utilize many features such as user interests, user authority and temporal features for this purpose (Pennacchiotti, Silvestri, Vahabi, & Venturini, 2012). However, since our main goal is not to propose a retweet prediction system, we adopt a simple algorithm which is only based on user interests. Given the tweets of two consecutive time intervals, i.e., November and December 2010, for a user u, we build her interest profile $P(u)$ based on the tweets that she has published or retweeted in November 2010. Further, we consider the tweets that she has retweeted in December 2010 as the ground truth to evaluate the results of the retweet prediction application. For user u, to predict a retweet, we consider the tweets of the users that she follows from whom she has retweeted at least one tweet in December 2010 as candidates, and compute the topic similarity between a candidate tweet and the user interest profile of user u as follows:

We represent each candidate tweet m as a weighted vector $P(m) = (f_m(z_1), \ldots, f_m(z_K))$ over the extracted topics $\mathbb{Z} = \{z_1, z_2, \ldots, z_K\}$. The value for $f_m(z_k)$ is calculated based on Eq. (10) (Wang et al., 2014):

$$f_m(z_k) = \frac{\prod_{c \in \mathbb{C}_m} w_{z_k}(c)}{\sum_{i=1}^{K} \prod_{c \in \mathbb{C}_m} w_{z_i}(c)} \tag{10}$$

where $\mathbb{C}_m$ denotes the set of concepts annotated in tweet m and $w_z(c)$ is the probability of concept c in topic $z \in \mathbb{Z}$.

Given $P(m)$ that represents a candidate tweet and $P(u)$, the interest profile of user u, it is possible to compute the interest of user u to tweet m by calculating the cosine similarity of the corresponding vectors of $P(m)$ and $P(u)$. We then 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, it is possible to evaluate the quality of retweet prediction, and therefore determine how successfully the interests of a user have been identified.

We evaluate the performance of the predictions using *Mean Average Precision (MAP)* and *nDCG at rank k (nDCG@k)* as two standard information retrieval metrics. The results are reported in Table 5 in terms of MAP, nDCG@5, nDCG@10 and nDCG@20. It can be observed that the two variants of our proposed approach EIUI, which are based on both users' explicit and implicit interests by utilizing social relations and topic relatedness outperform the EUI method which is solely based on textual content generated by the users. This means that incorporating implicit interests of users in addition to their explicit interests leads to user interest profiles that are more accurate for predicting relevant tweets to a given user. In other words, the content generated by users does not reveal

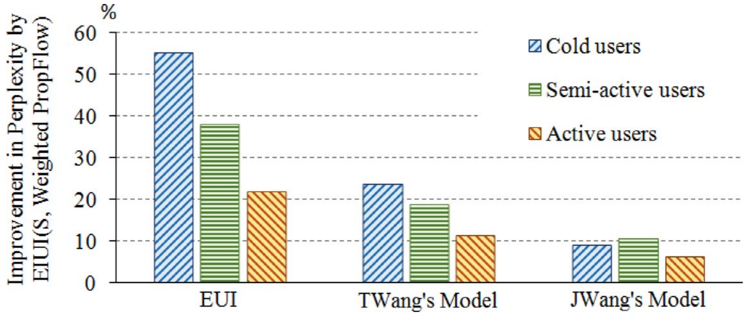

Fig. 7. Percentage of improvement of EIUI(*S*, Weighted PropFlow) over the baselines calculated based on the results reported in Table 4.

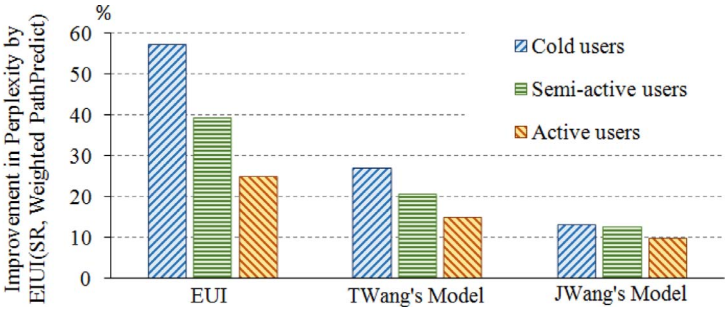

Fig. 8. Percentage of improvement of EIUI(*SR*, Weighted PathPredict) over the baselines calculated based on the results reported in Table 4.

Table 5
The performance of comparison methods in the context of a retweet prediction application in terms of MAP and NDCG. Based on a paired *t*-test, * indicates statistical significance.

Method	MAP	nDCG@5	nDCG@10	nDCG@20
EUI	.025	.026	.032	.036
TWang's Model	.026	.032	.035	.037
JWang's Model	.027	.035	.037	.038
EIUI (S, Weighted PropFlow)	.032*	.062	.059*	**.056***
EIUI (SR, Weighted PathPredict)	**.034***	**.069***	**.061***	**.056***

sufficient clues to extract all of the user's interests. Therefore, the incorporation of user's social interaction and the relatedness between active topics can lead to a more accurate representation of users' interests and consequently improve the quality of recommendations.

Based on the results in Table 5, our approach EIUI (*S*, Weighted PropFlow) that incorporates the relatedness between topics to infer implicit user interests offers superior results compared to both TWang's model (Wang et al., 2013) and JWang's model (Wang et al., 2014) which are based on the social connections between users and do not take into account the relationship between topics. This means that relatedness between topics is a strong clue to infer implicit interests of users and consequently results in user profiles that are more useful for retweet prediction compared to those that only consider user relations.

As another observation, by comparing two variants of our proposed framework, it can be seen that EIUI (*SR*, Weighted PathPredict) outperforms EIUI (*S*, Weighted PropFlow). This means, although semantic relatedness between topics is a better clue for inferring implicit interests of users compared to social relations, both of them are useful and our model can take advantage of both of them by designing a comprehensive representation model and applying a heterogeneous link prediction method on it.

We tested the statistical significance of the observed differences between the proposed method and each of the baselines by performing a paired *t*-test with 95% confidence level. As depicted in Table 5, the improvement of our approach EIUI (*SR*, Weighted PathPredict) is statistically significant over the baselines in terms of all the metrics.

Cold Start User Problem: We further analyze the different user interest detection strategies by investigating the impact of different level of user engagement on Twitter, on their performance. To do so, given the three user groups introduced in Section 4.3.2, we evaluate the performance of the predictions for each user groups, separately. The results in terms of MAP, nDCG@5, nDCG@10 and nDCG@20 are reported in Tables 6–9, respectively. As shown in these tables, in case of the two first user groups, our proposed approach EIUI (*SR*, Weighted PathPredict) significantly outperforms the baselines in terms of all the metrics. However, in case of active users, as it is highlighted in the tables (the bold values), the baseline methods give competitive or better performance compared to our approach EIUI. By performing a paired *t*-test between the results of our proposed approach and each of the baselines, we found that their improvement is not statistically significant for active users. This highlights that, compared to the baselines, our approach is more effective in the case of cold users and semi-active uses whose available tweet set is sparse; while showing competitive performance for the active users where the users have sufficient tweet content available.

Table 6
The performance of comparison methods in the context of a retweet prediction application for different user groups in terms of MAP. Based on a paired *t*-test, * indicates statistical significance.

Method	Cold users	Semi-active users	Active users
EUI	.012	.021	.053
TWang's Model	.013	.022	.054
JWang's Model	.013	.022	**.056**
EIUI (S, Weighted PropFlow)	.027*	.028*	.054
EIUI (SR, Weighted PathPredict)	**.03***	**.029***	**.056**

Table 7
The performance of comparison methods in the context of a retweet prediction application for different user groups in terms of nDCG@5. Based on a paired *t*-test, * indicates statistical significance.

Method	Cold users	Semi-active users	Active users
EUI	.026	.017	.076
TWang's Model	.026	.017	**.086**
JWang's Model	.026	.026	.076
EIUI (S, Weighted PropFlow)	.181	.040	.056
EIUI (SR, Weighted PathPredict)	**.191***	**.047***	.056

Table 8
The performance of comparison methods in the context of a retweet prediction application for different user groups in terms of nDCG@10. Based on a paired *t*-test, * indicates statistical significance.

Method	Cold users	Semi-active users	Active users
EUI	.022	.025	.074
TWang's Model	.022	.026	**.081**
JWang's Model	.023	.029	**.081**
EIUI (S, Weighted PropFlow)	.123	**.046***	.061
EIUI (SR, Weighted PathPredict)	**.134***	**.046***	.061

Table 9
The performance of comparison methods in the context of a retweet prediction application for different user groups in terms of nDCG@20. Based on a paired *t*-test, * indicates statistical significance.

Method	Cold users	Semi-active users	Active users
EUI	.025	.028	.078
TWang's Model	.025	.028	**.084**
JWang's Model	.025	.031	.078
EIUI (S, Weighted PropFlow)	.085	**.046***	.069
EIUI (SR, Weighted PathPredict)	**.096***	**.046***	.065

5. Discussion

The goal of our work has been to identify users' interests towards a set of active topics. Prior research has already shown evidence based on homophily theory that users' interests on social networks are not necessarily independent. Thus, pointing to the fact that it might be possible to infer a user's interests based on her social connections (Wang et al., 2013; Wen & Lin, 2010). Further, some researchers have argued that users often have coherent and related interests (Bhattacharya et al., 2014; Shen et al., 2013). Inspired from these insights, we modeled the problem of inferring implicit interests of users as a link prediction task over a representation model that includes three types of information: user explicit contributions to the topics, social connections between users, and the relatedness between the topics.

Our proposed representation model allows us to investigate the influence of both users' relationships and different measures of topic relatedness on the performance of the implicit user interest detection method. To do so, we followed two possible solutions for handling the link prediction problem: i.e., homogeneous and heterogeneous approaches and systematically explored different variants of our representation model by applying some well-known link prediction strategies. The results showed that the relatedness between topics is a more accurate clue for inferring implicit interests of users when compared to social relations, reinforcing the observation that users on Twitter are predominantly interested in semantically related topics.

Further, we found that by applying heterogeneous link prediction approach that distinguishes between different node and edge types of the graph, it is possible to take advantage of both topic relatedness and social relations simultaneously in our model. Another observation was that incorporating the strength of relationships in the representation model can contribute to the improvement of the performance of our implicit interest detection method.

After inferring implicit interests of users, our proposed work utilized both users' explicit and implicit interests to build their interest profile. To evaluate the predictive power of our proposed approach to infer unseen/implicit interests of users, we compared it with other state-of-the-art baselines. The results showed that our work is able to improve the state of the art in user interest detection methods on Twitter in terms of perplexity and in the context of a retweet prediction application. Based on the experiments, we found that the content generated by users does not reveal all possible clues to extract all of the user's interests. Therefore, the incorporation of user's social interaction and the relatedness between active topics can lead to a more accurate representation of users' interests. In other words, incorporating implicit interests of users in addition to their explicit interests leads to more accurate user interest profiles.

To investigate the user engagement level on Twitter on the performance of different user modeling strategies, we partitioned the users into three groups based on the number of posts they have published or retweeted. In case of active users who have published many tweets, all the methods offer a similar performance for inferring user interests. However, the results showed that our proposed approach show statistically significant better performance in case of cold users who do not have sufficient number of available tweets to be used for extracting their explicit interests. It is interesting to note that, most people are passive on twitter and prefer to passively read rather than to actively engage (Romero et al., 2011). In other words, as it is also depicted in Fig. 3 regarding our sample Twitter dataset, Twitter suffers from participation inequality, where a minority of users usually contribute the most while the others just free-ride. Therefore, inferring interests for passive users and consequently providing better recommendations for them will have impact on a higher number of users. This is the significant advantage of our method that is able to determine user interests for the cold and semi-active users, which is not something that is done effectively by the baselines.

6. Concluding rematks and future work

In this paper, we have proposed an approach for identifying user interests toward a set of topics on Twitter. We have modeled the problem of inferring implicit interests of users as a link prediction task over a graph representation model that includes three types of information: user explicit contributions to the topics, social connections between users, and the relatedness between the topics. In our evaluation, we first investigated the influence of different factors included in our representation model on the performance of the implicit user interest detection problem. Then, we compared our proposed approach for building user interest profiles with other state-of-the-art baselines in terms of perplexity and in the context of the retweet prediction application. The results showed that our work is able to significantly improve the state of the art. We further investigated the impact of our work on the three different categories of users classified as cold start, semi-active and active users. We were able to show that the highlight of our work is that it can significantly outperform the state of the art baselines in case of both cold start and semi-active users.

There are several directions, which we would like to explore in the future. Based on the fact that user interests change over time, we intend to include temporal behavior of users toward topics in our framework, and model interest evolution over time by considering the current interests of a user as a function of her interests in the previous time intervals. We intend to utilize the hierarchical knowledge of Wikipedia in our user interest detection framework which enables us to model user's high level interests more accurately and consequently can lead to improved quality of user interest predictions. As a proof of concept, in our recently published work (Zarrinkalam, Fani, Bagheri, & Kahani, 2017), we integrated the semantic information derived from the Wikipedia category structure and the temporal evolution of user's interests into our work to predict users future interests. As another future work, given that the characteristics of Twitter data such as user population and user tweeting behaviour may have changed over time (Liu, Kliman-Silver, & Mislove, 2014b), to investigate the influence of Twitter evolution on the performance of our model and the validity of our current findings, we intend to evaluate our proposed model on Twitter data, which are published more recently compared to the 2010 dataset that is used in this paper.

Acknowledgment

The author's acknowledge Hossein Fani's constructive contribution to the paper.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ipm.2017.12.003

References

Abel, F., Gao, Q., Houben, G., & Tao, K. (2011). *Analyzing user modeling on twitter for personalized news recommendations. User modeling, adaption and personalization - 19th international conference, UMAP 2011, Girona, Spain, July 11–15, 2011. proceedings*1–12. http://dx.doi.org/10.1007/978-3-642-22362-4_1.

Abel, F., Herder, E., Houben, G., Henze, N., & Krause, D. (2013). Cross-system user modeling and personalization on the social web. *User Modeling and User-Adapted Interaction, 23*(2–3), 169–209. http://dx.doi.org/10.1007/s11257-012-9131-2.

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social Networks, 25*(3), 211–230. http://dx.doi.org/10.1016/S0378-8733(03)00009-1.

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering, 17*(6), 734–749. http://dx.doi.org/10.1109/TKDE.2005.99.

Aiello, L. M., Petkos, G., Martín, C. J., Corney, D., Papadopoulos, S., Skraba, R., et al. (2013). Sensing trending topics in twitter. *IEEE Transactions on Multimedia, 15*(6), 1268–1282. http://dx.doi.org/10.1109/TMM.2013.2265080.

Alvarez-Melis, D., & Saveski, M. (2016). *Topic modeling in twitter: Aggregating tweets by conversations. Proceedings of the tenth international conference on web and social media, Cologne, Germany, May 17–20, 2016*519–522.

Bhattacharya, P., Zafar, M. B., Ganguly, N., Ghosh, S., & Gummadi, K. P. (2014). *Inferring user interests in the twitter social network. Eighth ACM conference on recommender systems, recsys '14, Foster City, Silicon Valley, Ca, USA - October 06 - 10, 2014*357–360. http://dx.doi.org/10.1145/2645710.2645765.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159. http://dx.doi.org/10.1016/S0031-3203(96)00142-2.

Calegari, S., & Pasi, G. (2013). Personal ontologies: Generation of user profiles based on the YAGO ontology. *Information Processing and Management, 49*(3), 640–658. http://dx.doi.org/10.1016/j.ipm.2012.07.010.

Cao, B., Kong, X., & Yu, P. S. (2014). *Collective prediction of multiple types of links in heterogeneous information networks. 2014 IEEE international conference on data mining, ICDM 2014, Shenzhen, China, December 14–17, 2014*50–59. http://dx.doi.org/10.1109/ICDM.2014.25.

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering, 26*(12), 2928–2941. http://dx.doi.org/10.1109/TKDE.2014.2313872.

Cornolti, M., Ferragina, P., & Ciaramita, M. (2013). *A framework for benchmarking entity-annotation systems. 22nd international world wide web conference, WWW '13, Rio de Janeiro, Brazil, May 13–17, 2013*249–260.

Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., et al. (2015). Analysis of named entity recognition and linking for tweets. *Information*

Processing and Management, 51(2), 32–49. http://dx.doi.org/10.1016/j.ipm.2014.10.006.

Duong, P. H., Nguyen, H. T., & Nguyen, V. P. (2016). *Evaluating semantic relatedness between concepts. Proceedings of the 10th international conference on ubiquitous information management and communication, IMCOM 2016, Danang, Vietnam, January 4–6, 2016*20:1–20:8.

Farzindar, A., & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence, 31*(1), 132–164. http://dx.doi.org/10.1111/coin.12017.

Feng, W., & Wang, J. (2013). *Retweet or not?: personalized tweet re-ranking. Sixth ACM international conference on web search and data mining, WSDM 2013, Rome, Italy, February 4–8, 2013*577–586. http://dx.doi.org/10.1145/2433396.2433470.

Ferragina, P., & Scaiella, U. (2012). Fast and accurate annotation of short texts with wikipedia pages. *IEEE Software, 29*(1), 70–75. http://dx.doi.org/10.1109/MS.2011.122.

He, W., Liu, H., He, J., Tang, S., & Du, X. (2015). *Extracting interest tags for non-famous users in social network. Proceedings of the 24th ACM international conference on information and knowledge management, CIKM 2015, Melbourne, vic, Australia, October 19 - 23, 2015*861–870. http://dx.doi.org/10.1145/2806416.2806514.

Huang, J., Peng, M., Wang, H., Cao, J., Gao, W., & Zhang, X. (2017). A probabilistic method for emerging topic tracking in microblog stream. *World Wide Web, 20*(2), 325–350. http://dx.doi.org/10.1007/s11280-016-0390-4.

Jiang, Y., Bai, W., Zhang, X., & Hu, J. (2017). Wikipedia-based information content and semantic similarity computation. *Information Processing and Management, 53*(1), 248–265. http://dx.doi.org/10.1016/j.ipm.2016.09.001.

Jiang, Y., Zhang, X., Tang, Y., & Nie, R. (2015). Feature-based approaches to semantic similarity assessment of concepts using wikipedia. *Information Processing and Management, 51*(3), 215–234. http://dx.doi.org/10.1016/j.ipm.2015.01.001.

Kabbur, S., Ning, X., & Karypis, G. (2013). *FISM: factored item similarity models for top-n recommender systems. The 19th ACM SIGKDD international conference on knowledge discovery and data mining, KDD 2013, Chicago, Il, USA, August 11–14, 2013*659–667. http://dx.doi.org/10.1145/2487575.2487589.

Kapanipathi, P., Jain, P., Venkatramani, C., & Sheth, A. P. (2014). *User interests identification on twitter using a hierarchical knowledge base. The semantic web: Trends and challenges - 11th international conference, ESWC 2014, Anissaras, Crete, Greece, May 25–29, 2014. proceedings*99–113. http://dx.doi.org/10.1007/978-3-319-07443-6_8.

Kapanipathi, P., Orlandi, F., Sheth, A. P., & Passant, A. (2011). *Personalized filtering of the twitter stream. Proceedings of the second workshop on semantic personalized information management: Retrieval and recommendation 2011, Bonn, Germany, October 24, 2011*6–13.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika, 18*(1), 39–43.

Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016). *Topic modeling for short texts with auxiliary word embeddings. Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval, SIGIR 2016, Pisa, Italy, July 17–21, 2016*165–174. http://dx.doi.org/10.1145/2911451.2911499.

Li, Y., Jiang, J., Liu, T., Qiu, M., & Sun, X. (2017). Personalized microtopic recommendation on microblogs. *ACM Transactions on Intelligent Systems and Technology, 8*(6), 77:1–77:21. http://dx.doi.org/10.1145/2932192.

Liben-Nowell, D., & Kleinberg, J. M. (2007). The link-prediction problem for social networks. *JASIST, 58*(7), 1019–1031. http://dx.doi.org/10.1002/asi.20591.

Lichtenwalter, R., & Chawla, N. V. (2011). Lpmade: Link prediction made easy. *Journal of Machine Learning Research, 12*, 2489–2492.

Lichtenwalter, R., Lussier, J. T., & Chawla, N. V. (2010). *New perspectives and methods in link prediction. Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, Washington, DC, USA, July 25–28, 2010*243–252. http://dx.doi.org/10.1145/1835804.1835837.

Liu, X., Yu, Y., Guo, C., & Sun, Y. (Yu, Guo, Sun, 2014a). *Meta-path-based ranking with pseudo relevance feedback on heterogeneous graph for citation recommendation. Proceedings of the 23rd ACM international conference on conference on information and knowledge management, CIKM 2014, Shanghai, China, November 3–7, 2014*121–130. http://dx.doi.org/10.1145/2661829.2661965.

Liu, Y., Kliman-Silver, C., & Mislove, A. (Kliman-Silver, Mislove, 2014b). *The tweets they are a-changin: Evolution of twitter users and behavior. Proceedings of the eighth international conference on weblogs and social media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1–4, 2014*.

Lu, L., & Zhou, T. (2009). *Role of weak ties in link prediction of complex networks. Proceeding of the ACM first international workshop on complex networks meet information & knowledge management, CIKM-CNIKM 2009, Hong Kong, China, November 6, 2009*55–58. http://dx.doi.org/10.1145/1651274.1651285.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 415–444.

Meguebli, Y., Kacimi, M., Doan, B., & Popineau, F. (2017). Towards better news article recommendation - with the help of user comments. *World Wide Web, 20*(6), 1293–1312. http://dx.doi.org/10.1007/s11280-017-0436-2.

Michelson, M., & Macskassy, S. A. (2010). *Discovering users' topics of interest on twitter: A first look. Proceedings of the fourth workshop on analytics for noisy unstructured text data, AND 2010, Toronto, Ontario, Canada, October 26th, 2010 (in conjunction with CIKM 2010)*73–80. http://dx.doi.org/10.1145/1871840.1871852.

Pennacchiotti, M., Silvestri, F., Vahabi, H., & Venturini, R. (2012). *Making your interests follow you on twitter. 21st ACM international conference on information and knowledge management, cikm'12, Maui, Hi, USA, October 29, - november 02, 2012*165–174. http://dx.doi.org/10.1145/2396761.2396786.

Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011). *Influence and passivity in social media. Machine learning and knowledge discovery in databases - european conference, ECML PKDD 2011, Athens, Greece, September 5–9, 2011, proceedings, part III*18–33. http://dx.doi.org/10.1007/978-3-642-23808-6_2.

de Sa, H. R., & Prudêncio, R. B. C. (2011). *Supervised link prediction in weighted networks. The 2011 international joint conference on neural networks, IJCNN 2011, San Jose, California, USA, July 31, - august 5, 2011*2281–2288. http://dx.doi.org/10.1109/IJCNN.2011.6033513.

Shen, W., Wang, J., Luo, P., & Wang, M. (2013). *Linking named entities in tweets with knowledge base via user interest modeling. The 19th ACM SIGKDD international conference on knowledge discovery and data mining, KDD 2013, Chicago, Il, USA, August 11–14, 2013*68–76. http://dx.doi.org/10.1145/2487575.2487686.

Shi, C., Li, Y., Zhang, J., Sun, Y., & Yu, P. S. (2017). A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering, 29*(1), 17–37. http://dx.doi.org/10.1109/TKDE.2016.2598561.

Spasojevic, N., Yan, J., Rao, A., & Bhattacharyya, P. (2014). *LASTA: large scale topic assignment on multiple social networks. The 20th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '14, New York, NY, USA - August 24, - 27, 2014*1809–1818. http://dx.doi.org/10.1145/2623330.2623350.

Srijith, P. K., Hepple, M., Bontcheva, K., & Preotiuc-Pietro, D. (2017). Sub-story detection in twitter with hierarchical dirichlet processes. *Information Processing and Management, 53*(4), 989–1003. http://dx.doi.org/10.1016/j.ipm.2016.10.004.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. L. (2004). *Probabilistic author-topic models for information discovery. Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, Seattle, Washington, USA, August 22–25, 2004*306–315. http://dx.doi.org/10.1145/1014052.1014087.

Sun, Y., Aggarwal, C. C., & Han, J. (2012). Relation strength-aware clustering of heterogeneous information networks with incomplete attributes. *PVLDB, 5*(5), 394–405.

Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C., & Han, J. (Barber, Gupta, Aggarwal, Han, 2011a). *Co-author relationship prediction in heterogeneous bibliographic networks. International conference on advances in social networks analysis and mining, ASONAM 2011, Kaohsiung, Taiwan, 25–27 july 2011*121–128. http://dx.doi.org/10.1109/ASONAM.2011.112.

Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. (Han, Yan, Yu, Wu, 2011b). Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB, 4*(11), 992–1003.

Wang, J., Zhao, W. X., He, Y., & Li, X. (2014). Infer user interests via link structure regularization. *ACM TIST, 5*(2), 23:1–23:22. http://dx.doi.org/10.1145/2499380.

Wang, T., Liu, H., He, J., & Du, X. (2013). *Mining user interests from information sharing behaviors in social media. Advances in knowledge discovery and data mining, 17th pacific-asia conference, PAKDD 2013, Gold Coast, Australia, April 14–17, 2013, proceedings, part II*85–98. http://dx.doi.org/10.1007/978-3-642-37456-2_8.

Welch, M. J., Schonfeld, U., He, D., & Cho, J. (2011). *Topical semantics of twitter links. Proceedings of the forth international conference on web search and web data mining, WSDM 2011, Hong Kong, China, February 9–12, 2011*327–336. http://dx.doi.org/10.1145/1935826.1935882.

Wen, Z., & Lin, C. (2010). *On the quality of inferring interests from social neighbors. Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, Washington, DC, USA, July 25–28, 2010*373–382. http://dx.doi.org/10.1145/1835804.1835853.

Weng, J., Lim, E., Jiang, J., & He, Q. (2010). *Twitterrank: finding topic-sensitive influential twitterers. Proceedings of the third international conference on web search and web data mining, WSDM 2010, New York, NY, USA, February 4–6, 2010*261–270. http://dx.doi.org/10.1145/1718487.1718520.

Xu, Z., Lu, R., Xiang, L., & Yang, Q. (2011). *Discovering user interest on twitter with a modified author-topic model. Proceedings of the 2011 IEEE/WIC/ACM international* *conference on web intelligence, WI 2011, campus scientifique de la doua, Lyon, France, August 22–27, 2011*422–429. http://dx.doi.org/10.1109/WI-IAT.2011.47.

Yang, L., Sun, T., Zhang, M., & Mei, Q. (2012). *We know what @you #tag: does the dual role affect hashtag adoption? Proceedings of the 21st world wide web conference 2012, WWW 2012, Lyon, France, April 16–20, 2012*261–270. http://dx.doi.org/10.1145/2187836.2187872.

Yu, Y., Wang, C., & Gao, Y. (2014). Attributes coupling based item enhanced matrix factorization technique for recommender systems. *CoRR, abs/1405.0770*.

Zarrinkalam, F., Fani, H., Bagheri, E., & Kahani, M. (2016). *Inferring implicit topical interests on twitter. Advances in information retrieval - 38th European conference on IR research, ECIR 2016, Padua, Italy, March 20–23, 2016. proceedings*479–491. http://dx.doi.org/10.1007/978-3-319-30671-1_35.

Zarrinkalam, F., Fani, H., Bagheri, E., & Kahani, M. (2017). *Predicting users' future interests on twitter. Advances in information retrieval - 39th European conference on IR research, ECIR 2017, Aberdeen, UK, April 8–13, 2017, proceedings*464–476. http://dx.doi.org/10.1007/978-3-319-56608-5_36.

Zarrinkalam, F., Fani, H., Bagheri, E., Kahani, M., & Du, W. (2015). *Semantics-enabled user interest detection from twitter. IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology, WI-IAT 2015, Singapore, December 6–9, 2015 - volume I*469–476. http://dx.doi.org/10.1109/WI-IAT.2015.182.