Inferring Implicit Topical Interests on Twitter

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

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

Abstract. Inferring user interests from their activities in the social network space has been an emerging research topic in the recent years. While much work is done towards detecting *explicit* interests of the users from their social posts, less work is dedicated to identifying *implicit* interests, which are also very important for building an accurate user model. In this paper, a graph based link prediction schema is proposed to infer implicit interests of the users towards emerging topics on Twitter. The underlying graph of our proposed work uses three types of information: user's followerships, user's explicit interests towards the topics, and the relatedness of the topics. To investigate the impact of each type of information on the accuracy of inferring user implicit interests, different variants of the underlying representation model are investigated along with several link prediction strategies in order to infer implicit interests. Our experimental results demonstrate that using topics relatedness information, especially when determined through semantic similarity measures, has considerable impact on improving the accuracy of user implicit interest prediction, compared to when followership information is only used.

Keywords: Implicit interest · Twitter · Topic relatedness · Collaborative filtering

1 Introduction

The growth of social networks such as Twitter has allowed users to share and publish posts on a variety of social events as they happen, in real time, even before they are released in traditional news outlets. This has recently attracted many researchers to analyze posts to understand the current emerging topics/events on Twitter in a given time interval by viewing each topic as a combination of temporally correlated words/terms or semantic concepts [2,4]. For instance, on 2 December 2010, Russia and Qatar were selected as the locations for the 2018 and 2022 FIFA World Cups. By looking at Twitter data on this

N. Ferro et al. (Eds.): ECIR 2016, LNCS 9626, pp. 479–491, 2016.
DOI: 10.1007/978-3-319-30671-1_35

day, a combination of keywords like *'FIFA World Cup'*, *'Qatar'*, *'England'* and *'Russia'* have logically formed a topic to represent this event.

The ability to model user interests towards these emerging topics provides the potential for improving the quality of the systems that work on the basis of user interests such as news recommender systems [21]. Most existing approaches build a user interest profile based on the explicit contribution of the user to the emerging topics [1,15]. However, such approaches struggle to identify a user's interests if the user has not explicitly talked about them. Consider the tweets posted by Mary:

- *"Qatar's bid to host the 2022 World Cup is gaining momentum, worrying the U.S., which had been the favorite* http://on.wsj.com/a8j3if"
- *"Russia rests 2018 World Cup bid on belief that big and bold is best | Owen Gibson (Guardian)* http://feedzil.la/g2Mpbs"

Based on the keywords explicitly mentioned by Mary in her tweets, one could easily infer that she is interested in the Russia and Qatar's selection as the hosts for the 2018 and 2022 FIFA World Cups. We refer to such interests that are directly derivable from a user's tweets as *explicit interests*. Expanding on this example, another topic emerged later in 2010, which was related to Prince William's engagement. Looking at Mary's tweets 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 never explicitly tweeted about the latter. If that is in fact the case, then Mary's user profile would need to include such an interest. We refer to these concealed user topical interests as *implicit interests*, i.e., topics that the user never explicitly engaged with but might have interest in.

The main objective of our work in this paper is to determine implicit interests of a user over the emerging topics in a given time interval. To this end, we propose to turn the implicit interest detection problem into a graph-based link prediction problem that operates over a heterogeneous graph by taking into account *(i)* users' interest profile built based on their explicit contribution towards the extracted topics, *(ii)* theory of Homophily [12], which refers to the tendency of users to connect to users with common interests or preferences; and *(iii)* relationship between emerging topics, based on their similar constituent contents and user contributions towards them. More specifically, the key contributions of our work are as follows:

- Based on the earlier works [7,21], we model users' interests over the emerging topic on Twitter through a set of correlated semantic concepts. Therefore, we are able to infer finer-grained implicit interests that refer to real-world events.
- We propose a graph-based framework to infer the implicit interests of users toward the identified topics through a link prediction strategy. Our work considers a heterogeneous graph that allows for including three types of information: user followerships, user explicit interests and topic relatedness.
- We perform extensive experimentation to determine the impact of one or a combination of these information types on accurately predicting the implicit

interests of users on Twitter, which provides significant insight on how users are explicitly and implicitly inclined towards emerging topics.

The rest of this paper is organized as follows. In the next section, we review the related work. Our framework to infer users' implicit interests is introduced in Sect. 3. Section 4 is dedicated to the details of our empirical experimentation and our findings. Finally, in Sect. 5, we conclude the paper.

2 Related Work

In this paper, we assume that an existing state of the art technique such as those proposed in [2,4] can be employed for extracting and modeling the emerging topics on Twitter as sets of temporally correlated terms/concepts. Therefore, we will not be engaged with the process of identification of the topics and will only focus on determining the implicit interest of users towards the topics once they are identified. Given this focus, we review the works that are related to the problem of user interest detection from social networks.

There are different works for extracting users' interests from social networks through the analysis of the users' generated textual content. Yang et al. [19] have modeled the user interests by representing her tweets as a bag of words, and by applying a cosine similarity measure to determine the similarity between the users in order to infer common interests. Xu et al. [18] have proposed an author-topic model where the latent variables are used to indicate if the tweet is related to the author's interests.

Since Bag of Words and Topic Modeling approaches are designed for normal length texts, they may not perform so effectively on short, noisy and informal text such as tweets. There are insufficient co-occurrence frequencies between keywords in short posts to enable the generation of appropriate word vector representations [5]. Furthermore, bag of words approaches overlook the underlying semantics of the text. To address these issues, some recent works have tried to utilize external knowledge bases to enrich the representation of short texts [8,13]. Abel et al. [1] have enriched Twitter posts by linking them to related news articles and then modeled user's interests by extracting the entities mentioned in the enriched messages. DBpedia and Freebase are often used for enriching Tweets by linking their content with unambiguous concepts from these external knowledge bases. Such an association provides explicit semantics for the content of a tweet and can hence be considered to be providing additional contextual information about the tweet [8,10]. The work in [21] has inferred fine grained user topics of interest by extracting temporally related concepts in a given time interval.

While most of the works mentioned above have focused on extracting explicit interests through analysing only textual contents of users, less work has been dedicated to inferring *implicit* interests of the users. Some authors have shown interest in the Homophily theory [12] to extract implicit interests. Based on this theory, users tend to connect to users with common interests or preferences.

Mislove et al. [14] have used this theory to infer missing interests of a user based on the information provided by her neighbors. Wang et al. [16] have extended this theory by extracting user interests based on implicit links between users in addition to explicit relations. While these works incorporate the relationship between users, they do not consider the relationship between the emerging topics themselves. In our work, we are interested to explore if a holistic view that considers the semantics of the topics, the user followership information and the explicit interests of users towards the topics can provide an efficient platform for identifying users' implicit interests.

In another line of work, semantic concepts and their relationships defined in external knowledge bases are leveraged to extract implicit user interests. Kapanipathi et al. [10] have extracted implicit interests of the user by mapping her primitive interests to the Wikipedia category hierarchy using a spreading activation algorithm. Similarly, Michelson and Macskassy [13] have identified the high-level interests of the user by traversing and analyzing the Wikipedia categories of entities extracted from the user's tweets. The main difference between the problem we tackle here from the previously mentioned works is that we view each topic of interest as a combination of correlated concepts as opposed to just a single concept. So the relationship between two topics is not predefined in the external knowledge base and we need to provide a measure of topic similarity or relatedness.

3 Implicit User Interest Prediction

The objective of our work is to model and identify implicit interests of a user, within a specific time interval T, towards the emerging topics on Twitter. To address this challenge, we propose to turn the implicit interest prediction problem into a link prediction problem that operates over a heterogeneous graph. We believe that in addition to user explicit contributions toward the emerging topics, there are two other types of information that can be considered to infer implicit interests of users, namely user followership relations and the possible relation between the emerging topics themselves. By considering this information as our representation model, the main research question we are seeking to answer in this paper is: 'which or what combination of these three types of information are most effective in allowing us to accurately identify a user's implicit interests?' Therefore, we propose a comprehensive graph-based representation model that includes these three types of information and is used in order to model the implicit interest identification problem.

3.1 Representation Model

Our underlying representation model can be formalized as follows:

Definition 1 (Representation Model). Let T be a specified time interval. Given a set of emerging topics and individual users at time interval T denoted by $\mathbb{Z}$ and U, respectively, our representation model $G = (G_U \cup G_{U\mathbb{Z}} \cup G_{\mathbb{Z}})$, is a heterogeneous graph composed of three subgraphs, G_U, $G_{U\mathbb{Z}}$ and $G_{\mathbb{Z}}$. $G_U = (V_U, E_U)$ is unweighted and directed, which represents followership relations between users on Twitter, $G_{U\mathbb{Z}} = (V_{U\mathbb{Z}}, E_{U\mathbb{Z}})$ represents explicitly observable user-topic relations and $G_{\mathbb{Z}} = (V_{\mathbb{Z}}, E_{\mathbb{Z}})$ denotes potential relationships between emerging topics in $\mathbb{Z}$.

In line with earlier work in the literature [1,21], we view each emerging topic $z \in \mathbb{Z}$ as a set of temporally correlated semantic concepts derived from an external knowledge base, i.e., Wikipedia, and model each topic in the following form:

Definition 2 (Emerging Topic). An emerging topic z at time interval T, is defined as a set of weighted semantic concepts $z = \{(c, w(c, z))|c \in C\}$, where $w(c, z)$ is a function that denotes the importance of concept c in topic z and C is the set of all semantic concepts observed at time interval T on Twitter.

In Definition 2, For instance, an emerging topic can be seen in our earlier example as a set $z_1 = \{$*'FIFA World Cup'*, *'Qatar'*, *'England'* and *'Russia'*$\}$, which is composed of four concepts from Wikipedia. Based on this topic representation model, the user-topic subgraph can be constructed based on the explicit mention of the topic by the user in her tweets.

Definition 3 (User-Topic Graph). A user-topic graph in time interval T, is a weighted directed graph $G_{U\mathbb{Z}} = (V_{U\mathbb{Z}}, E_{U\mathbb{Z}})$ where $V_{U\mathbb{Z}} = \mathbb{Z} \cup U$ and edges $E_{U\mathbb{Z}}$ are established by observing a user's explicit contributions towards any of the emerging topics. The weight of each edge $e_{uz} \in E_{U\mathbb{Z}}$ that ties user $u \in U$ to a topic $z \in \mathbb{Z}$ represents the degree of u's explicit interest in topic z in time interval T.

Our intuition for calculating the explicit interest of user $u \in U$ towards each topic z is that the more a user tweets about a certain topic, the more interested the user would be in that topic. We define the occurrence ratio of topic $z = \{(c, w(c, z))\}$ in tweet m, denoted $OR(z, m)$, as follows:

$$OR(z, m) = \frac{\sum_{c \in C} w(c, z) * \delta(c, m)}{\sum_{c \in C} w(c, z)} \tag{1}$$

where $\delta(c, m)$ is 1, if Tweet m is annotated with concept c, otherwise, $\delta(c, m) = 0$. The weight of each edge e_{uz} in $G_{U\mathbb{Z}}$ is calculated by averaging the value of $OR(z, m)$ over all tweets posted by the specific user u with regards to topic z.

Since we are interested in knowing whether potential relationships between topics can be used to infer implicit interests, the third type of information that we consider in our model is the relationship between the topics, i.e. topic-topic subgraph.

Definition 4 (Topic-Topic Graph). A topic-topic graph in time interval T, is a weighted undirected graph $G_{\mathbb{Z}} = (V_{\mathbb{Z}}, E_{\mathbb{Z}})$ where $V_{\mathbb{Z}}$ denotes the set of all emerging topics within time interval T, denoted by $\mathbb{Z}$, and $E_{\mathbb{Z}}$ denotes a set of edges representing the relationships between these topics. The weight of the edges between the topics in the topic-topic graph represents the degree of relatedness of the topics.

3.2 Topic Relatedness

There are three possible approaches through which the relation between the emerging topics can be identified in our model: *(i)* semantics relatedness, *(ii)* collaborative relatedness, and *(iii)* hybrid approach.

In the *semantic relatedness* approach, the relatedness of topics is determined based on the semantic similarity of their constituent concepts. In other words, two topics are considered to be similar if the concepts that make up the two topics are semantically similar. Given each topic in our model is composed of a set of Wikipedia concepts, the semantic relatedness of two emerging topics can be calculated by measuring the average pairwise semantic relatedness between the concepts of the two topics using a Wikipedia-based relatedness measure. In our experiments, we use WLM [17], which computes the concept relatedness through link structure analysis.

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 $G_{U\mathbb{Z}}$, we regard the problem of computing the collaborative relatedness of topics as an instance of a model-based collaborative filtering problem. To this end, we model the user-topic graph information as a user-item rating matrix R of size $|U| \times |\mathbb{Z}|$, 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 the user-topic graph $G_{U\mathbb{Z}}$, i.e., the degree of u's interest in topic 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 [9] that learns item-item similarities (topic relatedness) as a product of two rank matrices, P and Q. Two matrices P and Q denote latent factors of items. In our model, the rating for a given user u on topic z_i is estimated as:

$$\hat{r}_{ui} = b_u + b_i + (n_u^+)^{-\alpha} \sum_{j \in R_u^+} p_j q_i^T \tag{2}$$

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, 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 [24], matrices P and Q can be learnt by minimizing a regularized optimization problem:

$$minimize \frac{1}{2} \sum_{u,i \in R} ||r_{ui} - \hat{r}_{ui}||_F^2 + \frac{\beta}{2}(||P||_F^2 + ||Q||_F^2) + \frac{\lambda}{2}||b_u||_2^2 + \frac{\gamma}{2}||b_i||_2^2 \tag{3}$$

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 is computed as the dot product between the corresponding factors from P and Q i.e., p_i and q_j.

While the collaborative relatedness measure can find the topic relatedness based on the user's contributions to the topics, it overlooks the semantic relatedness between the two topics. 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 emerging topics. We follow the assumption of [20] for utilizing item attribute information to add the item relationship regularization term into Eq. (3). Based on this, 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:

$$\frac{\delta}{2}\sum_{i=1}^{|\mathbb{Z}|}\sum_{i'=1}^{|\mathbb{Z}|} S_{ii'}(||q_i - q_{i'}||_F^2 + ||p_i - p_{i'}||_F^2) \tag{4}$$

where δ is a parameter to control the impact of topic information, S is a matrix in which $S_{ii'}$ denotes the similarity between topics z_i and $z_{i'}$ based on their attributes. In our approach, attributes of topics are their constituent concepts and $S_{ii'}$ is calculated by measuring the semantic relatedness of two topics as introduced earlier.

3.3 Implicit Interest Prediction

After building the representation model G, our problem is to infer whether a user $u \in U$ is implicitly interested in topic $z \in \mathbb{Z}$ for cases when no explicit interest between u and z is observed in G. In other words, we are going to find missing links of $G_{U\mathbb{Z}}$ by adopting an unsupervised link prediction strategy over observed links in G.

Most of the unsupervised link prediction strategies either generate scores based on vertex neighborhoods or path information [11]. Vertex neighborhood methods are based on the idea that two vertices 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. All of these methods are based on a predictive score function for ranking links that are likely to occur. According to the experiments done in [11], there is no single superior method among existing work and their quality is dependent on the structure of the specific graph under study. Therefore, in our experiments, we exploit various well-known link prediction strategies for inferring implicit interests of a user. These strategies are introduced in Table 1.

4 Experiments

We perform our experimentation to answer the following research question: 'how and to what extent do the three types of information present in our representation model facilitate the identification of implicit user interests on Twitter?'.

Table 1. The five link prediction strategies chosen for user implicit interest prediction

Adamic/Adar	$score(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log\|\Gamma(z)\|}$ $\Gamma(x)$: the set of neighbors of vertex x
Common neighbors	$score(x, y) = \Gamma(x) \cap \Gamma(y)$
Jaccard's coefficient	$score(x, y) = \|\Gamma(x) \cap \Gamma(y)\|/\|\Gamma(x) \cup \Gamma(y)\|$
Katz	$score(x, y) = \sum_{\ell=1}^{\infty} \beta^{\ell} \|path_{x,y}^{<\ell>}\|$ $\|path_{x,y}^{<\ell>}\|$: a set of all paths with length ℓ from x to y β: damping factor to give the shorter paths more weights
SimRank	$score(x, y) = sim(x, y)$ $sim(x, y) = \lambda(\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} sim(a, b))/\|\Gamma(x)\|\|\Gamma(y)\|$ $\lambda \in [0, 1] and sim(x, x) = 1$

4.1 Experimental Setup

Dataset. Our experiments were conducted on the available Twitter dataset presented by Abel et al. [1]. It consists of approximately 3M tweets sampled between November 1 and December 31, 2010. Since we needed followership information to build the user-user graph, we used the Twitter RESTful API to crawl these relationships.

Evaluation Methodology and Metrics. Our evaluation strategy is based on the *leave-one-out method*. At each time, we divide our representation model into a training set and a test set by randomly picking one pair <user, topic> from user-topic graph $G_{U\mathbb{Z}}$ for test and the rest of the representation model for training. We repeat this procedure for all pairs. To evaluate the results, we use two metrics: the Area Under Receiver Operating Characteristic (AUROC) and the Area Under the Precision-Recall (AUPR) curves [6].

Parameter Setting. In Topic detection step, we follow the approach proposed in [7] to extract the emerging topics ($\mathbb{Z}$) within a given time interval T. After detecting $\mathbb{Z}$, based on Definition 2, we need to compute the weight of each concept c in each topic z, i.e., $w(c, z)$. To do so, we utilize the Degree Centrality of concept vertex c in topic z computed by summing the weights attached to the edges connected to c in topic z [21]. Further, in the learning step of computing the collaborative relatedness between topics, we use the default parameter settings of the Librec library and set $\beta = \lambda = \gamma = \delta = 0.001$. The learning rate is set to 0.01, the number of item latent factors is set to 10 and the number of iterations to 100.

4.2 Results and Discussion

To answer our research question, we conduct a set of experiments in which different link prediction strategies are applied on variants of our representation model. There are two main variation points which are incorporated in our representation model: *(i)* followership information (F) and *(ii)* the type of topics relatedness measure, i.e., semantic (S), collaborative (C) or hybrid (CS). By selecting and combining the different alternatives, we obtain 7 variants that we will systematically compare in this section. We include user's explicit interest information in all of the seven variants. As some brief example on how to interpret the models, Model F only uses user followership information in addition to users' explicit interests. The SF Model considers topic relationships computed using semantic relatedness in addition to user followership and user's explicit interests. The rest of the models can be interpreted similarly.

In order to make a fair comparison, we repeat the experimentation for all the selected link prediction strategies introduced in Table 1. The results in terms of AUROC and AUPR are reported in Table 2. Given AUROC and AUPR values can be misleading in some cases, we also visually inspect the ROC curves in addition to the area under the curve values. Due to space limitation and also the elaborate theorem proved in [6] that a curve dominates in ROC space if and only if it dominates in PR space, we only present the ROC curves in Fig. 1.

As illustrated in Table 2 and Fig. 1, we can clearly see that the SimRank link prediction method has not shown a good performance over none of the variants. Based on our results, SimRank acts as a random predictor because for most of the models its AUROC value is about 0.5 and its ROC curve is near y=x. Therefore, in the rest of this section, to investigate the influence of the different variants of our representation model on the performance of inferring implicit interests of users we ignore the results of the SimRank strategy.

Table 2. The AUROC/AUPR values showing the performance of different model variants

Model	Metric	Adamic/ Adar	Common neighbor	Jaccard coefficient	Katz $\beta = 0.0005$	Katz $\beta = 0.005$	Katz $\beta = 0.5$	SimRank $\lambda = 0.8$
F	AUROC	0.500	0.500	0.500	0.524	0.524	0.528	0.510
	AUPR	0.438	0.438	0.438	0.454	0.454	0.458	0.422
S	AUROC	**0.791**	0.790	0.774	0.790	0.790	0.788	0.500
	AUPR	**0.740**	0.739	0.723	0.740	0.739	0.734	0.438
SF	AUROC	**0.791**	0.790	0.762	0.757	0.753	0.720	0.520
	AUPR	**0.740**	0.739	0.707	0.660	0.652	0.602	0.430
C	AUROC	0.712	0.710	0.700	0.714	0.715	0.728	0.500
	AUPR	0.657	0.651	0.610	0.657	0.661	0.680	0.438
CF	AUROC	0.773	0.771	0.758	0.742	0.738	0.716	0.517
	AUPR	0.717	0.714	0.692	0.647	0.640	0.602	0.428
CS	AUROC	0.762	0.761	0.748	0.763	0.763	0.767	0.500
	AUPR	0.697	0.695	0.661	0.699	0.699	0.707	0.438
CSF	AUROC	0.762	0.761	0.738	0.736	0.732	0.707	0.520
	AUPR	0.697	0.695	0.652	0.640	0.632	0.595	0.428

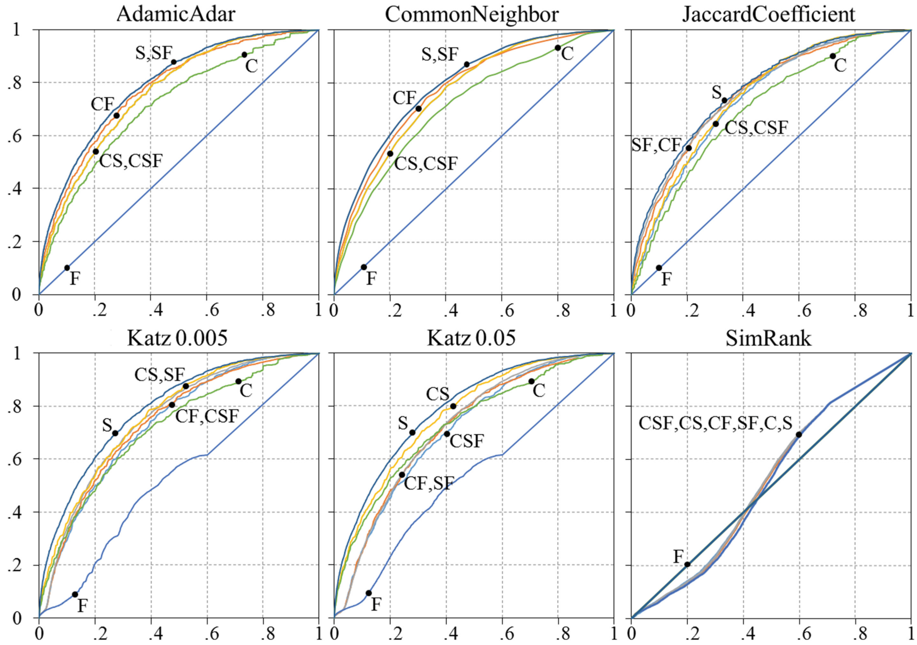

Fig. 1. The ROC curves for comparing the seven variants.

As mentioned earlier, Model F only considers followership information in addition to users' explicit interests to infer users' implicit interests. Instead, the models S, C and CS employ three different techniques for identifying topic relationships: model S uses semantic relatedness of the concepts included in the topics, model C uses collaborative relatedness and, model CS follows a hybrid approach. As depicted in Table 2, all these three models outperform Model F noticeably in terms of AUROC and AUPR. We can also see that the models S, C and CS dominate Model F in ROC space. This means that considering the relationships between the topics considerably improves the accuracy of inferring implicit interests in comparison with when only followership information is used.

By comparing S, C and CS themselves, it can be observed that using the semantic relatedness variant results in higher accuracy for the prediction of implicit interests compared to the collaborative and hybrid measures. This is an interesting observation that implies that users are predominantly interested in topics that are around similar topics. The three pairs of topics with the most relatedness obtained by the S model are shown in Fig. 2 (right). For an instance, the topics $z_1 = \{$*Chelsea F.C., Arsenal F.C.*$\}$ and $z_2 = \{$*FC Barcelona, Real Madrid C.F.*$\}$ refer to two derbies correspondingly in England and Spain. As confirmed by Wikipedia, these two competitions are among the most famous derbies in their countries and also in the world. As a result, it is reasonable to infer, with some lesser probability, that a user who is explicitly interested in one of these derbies, is probably interested also in the other one.

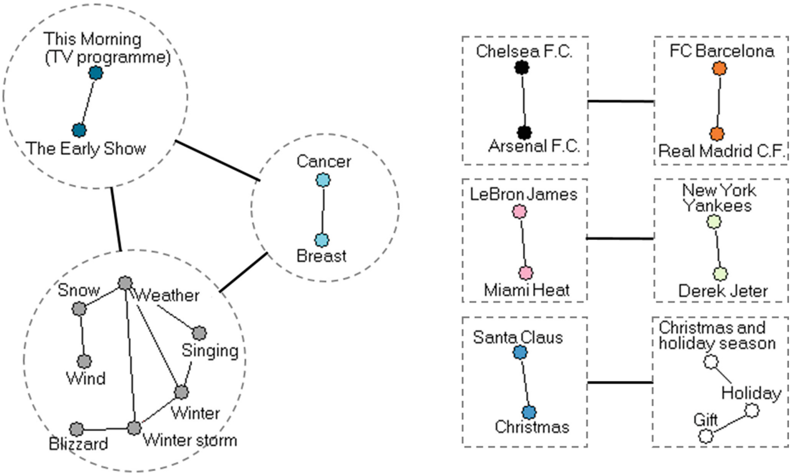

Fig. 2. Topmost related topics based on Hybrid (left) and semantic (right) measures

When looking at the results in Table 2, one can see that model C shows slightly weaker results compared to S, which can be the sign of two points: *(i)* semantic relatedness of topics is a more accurate indication of the tendency of users towards topics compared to collaborative relatedness of topics, and *(ii)* while C shows a weaker performance, its performance is in most cases only slightly weaker. This could mean that there is some degree of similarity between the results obtained by the two methods (C and S) pointing to the fact that even when using the collaborative relatedness measure, a comparable result to when the semantic relatedness measure is used can be obtained. Our explanation for this is that Twitter users seem to follow topics that are from similar domains or genres. This is an observation that is also reported in [3] and can be seen in the *Who Likes What* system. Therefore, when trying to predict a user's implicit interest, it would be logical to identify those that are on topics closely related to the user's explicit interests. Given this observation, the user's that are most similar within the context of collaborative filtering, are likely to also be following a coherent set of topics (not a variety of topics) and therefore, provide grounds for a reasonable estimation of the implicit interests.

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 semantic similarity of the concepts in each pair of topics, which is quite an inexpensive operation, whereas the computation of C and CS require solving an optimization problem through Stochastic Gradient Descent. Additionally, 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. As an example, the three top-most similar topics obtained by CS are illustrated in Fig. 2 (left). The topic $z_3 = \{$*The Early Show, This Morning*$\}$ refers to two popular TV programmes, the other one is related to weather forecasting and the last one focuses on breast cancer. It is clear that these topics are not semantically related to each other, however, the users who are explicitly interested in the two programmes are probably interested in knowing the weather forecast which is reported in these programmes. Further, the third topic shows that breast cancer was most likely a contentious hot topic on these two programmes in that time period; therefore, the user who followed the programmes also tweeted about this topics. While the topic connections between z_3 and weather and also breast cancer is logical, it would be a stretch to say those who are interested in breast cancer are also interested in knowing about the weather, and this is why the collaborative approach shows weaker results compared to the semantic approach.

As another observation, the models SF, CF and CSF incorporate the followership information correspondingly in the S, C and CS models. As demonstrated in Table 2, no uniform observation can be made in any of the cases, i.e., the followership information does not seem to have a noticeable impact on the results. As a result, through our experiments we were not able to show the impact of homophily theory that suggests the user interests can be extracted from their relationship to other users. In summary, model S, which relies solely on the semantic relatedness of topics and user's explicit contributions to these topics shows the best performance across all seven variants. The SF model shows the same performance as S in which the additional followership information does not seem to have impacted the final results.

5 Conclusions and Future Work

In this paper, we studied the problem of inferring implicit interests of a user toward a set of emerging topics on Twitter. We model this problem as a link prediction task over a graph including three type of information: followerships, users explicit interests and topic relatedness. To investigate the influence of different types of information on the performance of the implicit interest detection problem, we proposed different variants of our representation model and applied some well-known link prediction strategies. The results showed that considering the relationships between the topics considerably improves the accuracy compared to using only followership information. Further, it was our observation that users on Twitter are predominantly interested in the coherent and semantically related topics and not on unrelated topics. As future work, we are investigating meta-path-based relationship prediction framework for heterogeneous graphs as our link prediction strategy. Further, based on the idea that user interests change over time, we intend to include temporal behavior of users toward topics in our implicit user interest identification problem.

References

1. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Analyzing user modeling on twitter for personalized news recommendations. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) UMAP 2011. LNCS, vol. 6787, pp. 1–12. Springer, Heidelberg (2011)
2. Aiello, L.M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Goker, A., Kompatsiaris, I., Jaimes, A.: Sensing trending topics in twitter. IEEE Trans. Multimedia **15**(6), 1268–1282 (2013)
3. Bhattacharya, P., Muhammad, B.Z., Ganguly, N., Ghosh, S., Gummadi, K.P.: Inferring user interests in the twitter social network. In: RecSys 2014, pp. 357–360 (2014)
4. Cataldi, M., Di Caro, L., Schifanella, C.: Emerging topic detection on twitter based on temporal and social terms evaluation. In: IMDMKDD 2010, p. 4 (2010)
5. Cheng, X., Yan, X., Lan, Y., Guo, J.: Btm: Topic modeling over short texts. IEEE Trans. Knowl. Data Eng. **26**(12), 2928–2941 (2014)
6. Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: ICML 2006, pp. 233–240 (2006)
7. Fani, H., Zarrinkalam, F., Zhao, X., Feng, Y., Bagheri, E., Du, W.: Temporal identification of latent communities on twitter (2015). arXiv preprint arxiv:1509.04227
8. Ferragina, P., Scaiella, U.: Fast and accurate annotation of short texts with wikipedia pages. IEEE Softw. **29**(1), 70–75 (2012)
9. Kabbur, S., Ning, X., Karypis, G.: FISM: factored item similarity models for top-N recommender systems. In: KDD 2013, pp. 659–667 (2013)
10. Kapanipathi, P., Jain, P., Venkataramani, C., Sheth, A.: User interests identification on twitter using a hierarchical knowledge base. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8465, pp. 99–113. Springer, Heidelberg (2014)
11. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. Am. Soc. Inform. Sci. Technol. **58**(7), 1019–1031 (2007)
12. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Ann. Rev. Sociol. 27, pp. 415–444 (2001)
13. Michelson, M., Macskassy, S.A.: Discovering users' topics of interest on twitter: A first look. In: AND 2010, pp. 73–80 (2010)
14. Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: You are who you know: inferring user profiles in online social networks. In: WSDM 2010, pp. 251–260 (2010)
15. Shin, Y., Ryo, C., Park, J.: Automatic extraction of persistent topics from social text streams. World Wide Web **17**(6), 1395–1420 (2014)
16. Wang, J., Zhao, W.X., He, Y., Li, X.: Infer user interests via link structure regularization. TIST **5**(2), 23 (2014)
17. Witten, I., Milne, D.: An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In: WikiAI 2008, pp. 25–30 (2008)
18. Xu, Z., Lu, R., Xiang, L., Yang, Q.: Discovering user interest on twitter with a modified author-topic model. In: WI-IAT 2011, vol. 1, pp. 422–429 (2011)
19. Yang, L., Sun, T., Zhang, M., Mei, Q.: We know what@ you# tag: does the dual role affect hashtag adoption? In: WWW 2012, pp. 261–270 (2012)
20. Yu, Y., Wang, C., Gao, Y.: Attributes coupling based item enhanced matrix factorization technique for recommender systems. arXiv preprint. (2014). arxiv:1405.0770
21. Zarrinkalam, F., Fani, H., Bagheri, E., Kahani, M., Du, W.: Semantics-enabled user interest detection from twitter. In: WI-IAT 2015 (2015)