* This paper has been accepted at Hypertext 2017, final version can be found at ACM DL

# Leveraging Followee List Memberships for Inferring User Interests for Passive Users on Twitter


Guangyuan Piao
Insight Centre for Data Analytics
National University of Ireland, Galway
IDA Business Park
Galway, Ireland
guangyuan.piao@insight-centre.org

John G. Breslin
Insight Centre for Data Analytics
National University of Ireland, Galway
IDA Business Park
Galway, Ireland
john.breslin@nuigalway.ie



## ABSTRACT

User modeling for inferring user interests from Online Social Networks (OSNs) such as Twitter has received great attention in the user modeling community with the growing popularity of OSNs. The focus of previous works has been on analyzing user-generated content such as tweets to infer user interests. Therefore, these previous studies were limited to *active users* who have been actively generating content. On the other hand, with the percentage of passive use of OSNs on the rise, some researchers investigated different types of information about followees (i.e., people that a user is following) such as *tweets*, *usernames*, and *biographies* to infer user interests for *passive users* who use OSNs for consuming information from followees but who do not produce any content. Although different types of information about followees have been exploited, *list memberships* (a topical list which other Twitter users can freely add a user into) of followees have not yet been investigated extensively for inferring user interests.

In this paper, we investigate *list memberships* of followees, to infer interest profiles for passive users. To this end, we propose user modeling strategies with two different weighting schemes as well as a refined interest propagation strategy based on previous work. In addition, we investigate whether the information from *biographies* and *list memberships* of followees can complement each other, and thus improve the quality of inferred interest profiles for passive users. Results show that leveraging *list memberships* of followees is useful for inferring user interests when the number of followees is relatively small compared to using *biographies* of followees. In addition, we found that combining the two different types of information (*list memberships* and *biographies*) of followees can improve the quality of user interest profiles significantly compared to a state-of-art method in the context of link recommendations on Twitter.



## CCS CONCEPTS

**•Information systems → Personalization; Social recommendation;**

## KEYWORDS

User modeling; Personalization; Twitter; Passive users;

**ACM Reference format:**
Guangyuan Piao and John G. Breslin. 2016. Leveraging Followee List Memberships for Inferring User Interests for Passive Users on Twitter. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 10 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn


## 1 INTRODUCTION

With the rapid growth of Online Social Networks (OSNs), people can now consume rich, diverse information that previously was not available. According to a survey, one in three Web users seeks medical information using OSNs, and over 50 percent of users consume news in OSNs [27]. On the other hand, the huge volume of user-generated content causes an information overload problem for users consuming relevant information that they might be interested in. It has been reported that users follow 80 people on average on Twitter[1] [26], which results in hundreds or even thousands of tweets posted to each user every day. In this regard, it is important to infer user interest profiles based on user activities in OSNs such as Twitter to support personalized recommendations for content. Researchers have focused on *active users* who actively generate content on Twitter, and addressed the problem by exploiting user-generated content such as tweets to build user interest profiles. However, there is also an increasing number of *passive users* in OSNs[2]. For example, 44% of Twitter users have never sent a tweet according to a research done by Twopcharts[3]. Therefore, it is important to infer user interest profiles for those *passive users* who are only consuming information on Twitter and not generating any content. To this end, different types of information such as *tweets*, *usernames*, and *biographies* of followees have been exploited to infer user interest profiles for *passive users* on Twitter. ***Biographies*** (***bios***) on Twitter are *self-descriptions* of users, and it has been shown that exploiting *bios* of followees can provide improved user interest profiles of passive users compared to exploiting usernames or tweets of followees in a recent study [25]. For example, we can assume a user might be interested in "*Pokémon Go*" if the user is following another user who describes himself/herself as a "*Pokémon Go player*" in his/her biography on Twitter. In this paper, we investigate another type of information - ***list memberships*** of followees to infer user interests for passive users. *List memberships*



*Conference'17, Washington, DC, USA*



DOI: 10.1145/nnnnnnn.nnnnnnn


[1] https://twitter.com/
[2] http://www.corporate-eye.com/main/facebooks-growing-problem-passive-users/
[3] http://guardianlv.com/2014/04/twitter-users-are-not-tweeting/

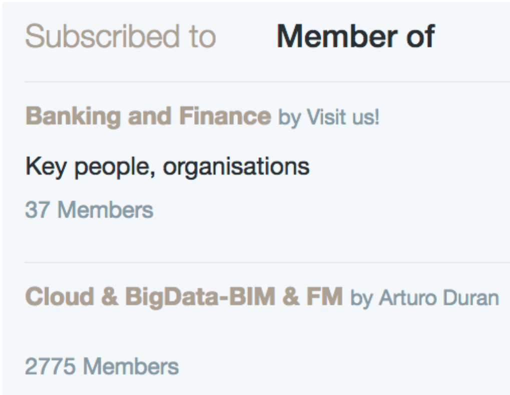



**Figure 1: An example of list memberships for a Twitter user.**

for a user on Twitter denote a topical list which the user has been added into by the list owners. Figure 1 shows an example of some *list memberships* that a Twitter user *@alice* has been added to by other users on Twitter. Different from *bios* (*self-descriptions*), *list memberships* can be seen as *others-descriptions* about *@alice*, which provide some third-party indications about what kind of topics *@alice* has been tweeting about on Twitter.

In this paper, we first propose a user modeling strategy leveraging *list memberships* of followees. In addition, we also explore whether the two different views (*self-descriptions* and *others-descriptions*) of followees can complement each other to improve the quality of inferred user interest profiles for *passive users* in the context of a link recommender system on Twitter. The contributions of this work are summarized as follows.

- We investigate whether *list memberships* of followees can provide sufficient and qualitative information for inferring user interests for *passive users* by applying two different weighting schemes and a refined interest propagation strategy.
- We combine the two different views (*self-descriptions* and *others-descriptions*) of followees to infer user interest profiles for *passive users* to study the synergetic effect of combining the two views.
- We evaluate our user modeling strategies for *passive users* in the context of link recommendations on Twitter compared to a state-of-art approach.

The organization of the rest of the paper is as follows. In Section 2, we give some related work. Section 3 describes our user modeling strategy which leverages the list memberships of users' followees to infer user interest profiles. In section 4, we present the experimental setup for our study. The results of our experiment are presented in Sections 5 and 6. Finally, Section 7 concludes the paper with some future work.

## 2 RELATED WORK

The first and fundamental step for user modeling is the representation of user interests. In order to represent user interest profiles, various approaches have been proposed in the literature [2, 3, 11, 12, 15, 18–20]. For example, Mislove et al. [18] proposed using *Bag of Words*, and Harvey et al. [12] proposed using a *Topic Modeling* approach to represent user interest profiles. Some previous studies also explored list memberships to build word-based user profiles. Kim et al. [15] explored the tweets published by the users in the same list to model the characteristics of the target user. Hannon et al. [11] exploited human-annotated tags of list memberships from third-party services such as Listorious to construct user interest profiles. However, these approaches focused on words, and the semantic information and relationships among words cannot be incorporated. Furthermore, the *Topic Modeling* approach based on the assumption that a single document contains rich information, which is not the case on Twitter. Some previous studies have shown that *Topic Modeling* approaches did not work well on Twitter [14, 17, 29].

To overcome the limitation of *word-based* approaches, researchers proposed using *Bag of Concepts* to represent user interest profiles. Here, a concept denotes an entity such as `Steve_Jobs` or a corresponding category of the entity such as `Apple_Inc._executives` based on the background knowledge from a knowledge base such as DBpedia [16]. For example, Abel et al. [2] compared three different representations of user interest profiles, and found that *entity-based* user profiles outperform *hashtag-* and *topic-based* user profiles on Twitter in the context of news recommendations. Some previous studies further exploited background knowledge linked to the concepts to enrich user interest profiles [14, 20, 21] with the *Bag of Concepts* approach e.g., using Wikipedia[4] entities or categories for representing user interests. For instance, Siehndel et al. [28] proposed constructing user interest profiles leveraging 23 top-level Wikipedia categories, which linked from the extracted entities from the tweets of a target user. Similarly, Kapanipathi et al. [14] first extracted Wikipedia entities from the tweets of a target user, and set those entities as activated nodes. Afterwards, they applied various spreading activation functions by exploiting refined Wikipedia categories to build *category-based* user interest profiles. Different from using Wikipedia categories, DBpedia has been used for propagating user interest profiles in some recent studies [20, 22] as it provides rich background knowledge about entities (e.g., related entities via different properties in addition to the categories of them). For example, Piao et al. [22] showed that considering different structures of background knowledge, i.e., categories and related entities, can improve the quality of user modeling on Twitter compared to exploiting categories only.

On top of a fixed representation of user interests, there are also some works studying temporal dynamics of user interests on Twitter based on the hypothesis that the interests of users change over time [1, 2, 5, 8, 20, 22, 23], which is not the focus on our work. In this study, we also use the *Bag of Concepts* approach for representing user interests, and use DBpedia as our background knowledge base. Although those previous works presented interesting results on user modeling in OSNs, most of them focused on *active users* who actively post tweets, to infer user interest profiles by analyzing users' tweets. Our work differs in that we focus on *passive users* who do not generate content on Twitter, but keep following other users to receive information they might be interested in.

[4] http://www.wikipedia.org

Some authors from previous studies [6, 7, 25, 28] also pointed out the needs to investigating other types of information beyond tweets for inferring user interest profiles. This line of work focuses on inferring interests for *passive users* who do not generate content (tweets), but mostly consume content from their followees. For example, Faralli et al. [10] and Besel et al. [6] proposed linking followees' accounts to Wikipedia entities based on followees' full names, and then propagate user interests leveraging Wikipedia categories. For instance, the entity `Cristiano_Ronaldo` would be found as a user's interest if the user was following `Cristiano_Ronaldo` on Twitter. Afterwards, the corresponding Wikipedia categories of the entity were leveraged to construct *category-based* user interest profiles by applying different propagation strategies. Faralli et al. [10] pointed out that the user interest profiles built by leveraging followee profiles are more stable and scalable compared to analyzing the tweets of followees. However, they also pointed out that linking Twitter accounts to Wikipedia entities is limited to a small percentage of famous users such as celebrities (e.g., less than 13% of followees can be linked to Wikipedia entities in [10]). To overcome this, the authors from [25] studied whether the *biographies* of followees can be exploited to provide useful information for user modeling for *passive users*. To this end, they fetched all of the followees of a target user first and then extracted DBpedia entities from the biographies of those followees. For example, the entity `Pokémon_GO` can be extracted based on a followee's bio "*Pokémon Go player*". Afterwards, the extracted entities were further used for propagating user interests based on background knowledge from DBpedia by exploiting their related entities as well as corresponding categories. The results from [25] showed that exploiting biographies of followees can provide quantitative and qualitative information for inferring user interests for passive users compared to the approach linking followees' accounts to Wikipedia entities. In this regard, we use this approach [25] as our baseline for evaluating our proposed user modeling strategies.

## 3 USER MODELING LEVERAGING LIST MEMBERSHIPS OF FOLLOWEES

In the same way as previous studies, user interest profiles in this work are represented using DBpedia concepts and corresponding weights. We use the same definition from [22] as follows.

*Definition 3.1.* The interest profile of a user $u \in U$ is a set of weighted DBpedia concepts. The weight with respect to $u$ for a concept $c \in C$ is computed by a weighting scheme $w(u, c)$.

$$P_u = \left\{ \left(c, w\left(u, c\right)\right) \mid c \in C \right\} \tag{1}$$

Here, $C$ and $U$ denote the set of concepts in DBpedia and set of users respectively. Concepts can be either *entities* or *categories* in DBpedia.

The general process of building user interest profiles based on *list memberships* of followees is shown in Figure 2. Given a Twitter user, we go through five main steps to construct an interest profile for the user.

(1) Fetch all of the user's followees.
(2) Fetch all *list memberships* of followees.
(3) Extract DBpedia entities from the *list memberships*.
(4) Construct *primary interests* based on the extracted entities by applying a weighting scheme.
(5) Apply an interest propagation strategy to *primary interests*.

First, for a given user $u$, the followees of $u$ and their *list memberships* can be fetched (steps 1 and 2) using the Twitter API[5]. Afterwards, DBpedia entities are extracted using the TAG.ME API[6] based on the full names of *list memberships*. For example, entities such as `Middle_East` and `Celebrity` can be extracted from *list memberships* with full names "*Middle East*" and "*Celebs*". Afterwards, these extracted entities are used to construct $u$'s ***primary interests***. Although the Aylien API[7] has been used for extracting entities for tweets and news articles in the literature [23, 25], we found the Aylien API is not the optimal choice for extracting entities from the names of *list memberships* due to the short nature of those names. Therefore, we use the TAG.ME API instead for extracting entities from *list memberships*.

### 3.1 Constructing Primary Interests

In this subsection, we introduce two different weighting schemes for weighting extracted entities in order to construct a user's *primary interests*.

- **Weighting Scheme 1 (WS1).** The intuitive way of weighting extracted entities from *list memberships* of followees is based on the the number of occurrences of these entities. However, directly summing the number of occurrences might be biased by followees who have a great number of *list memberships*. In this regard, we use a normalized sum of occurrences of entities from followees as a weighting scheme for constructing the *primary interests* of a target user $u$. For example, an interest profile of a followee $f \in F_u$ can be normalized as follows.

$$P_f = \left\{ \left(c, w\left(f, c\right)\right) \mid c \in C \right\} \tag{2}$$

where $\sum_{c_i \in C} w(f, c_i) = 1$, and $F_u$ denotes all the followees of a user $u$. Finally, the weight of an entity $c_j$ with respect to $u$ is measured as below:

$$w(u, c_j) = \sum_{f \in F_u} w(f, c_j). \tag{3}$$

- **Weighting Scheme 2 (WS2).** For a target user $u$, Chen et al. [9] aggregated the weight of each word from followees' tweets by excluding the words mentioned only in a single followee. Similarly, we aggregate the weight of each entity from followees' *list memberships* by excluding entities extracted only in a single followee. The weight of each entity in $u$'s profile $w(u, c_j)$ is calculated as $w(u, c_j) =$ *the number of followees who have $c_j$ in their list memberships.* Note that this weighting scheme does not care about the number of

[5] https://dev.twitter.com/
[6] https://tagme.d4science.org/tagme/
[7] http://aylien.com/text-api

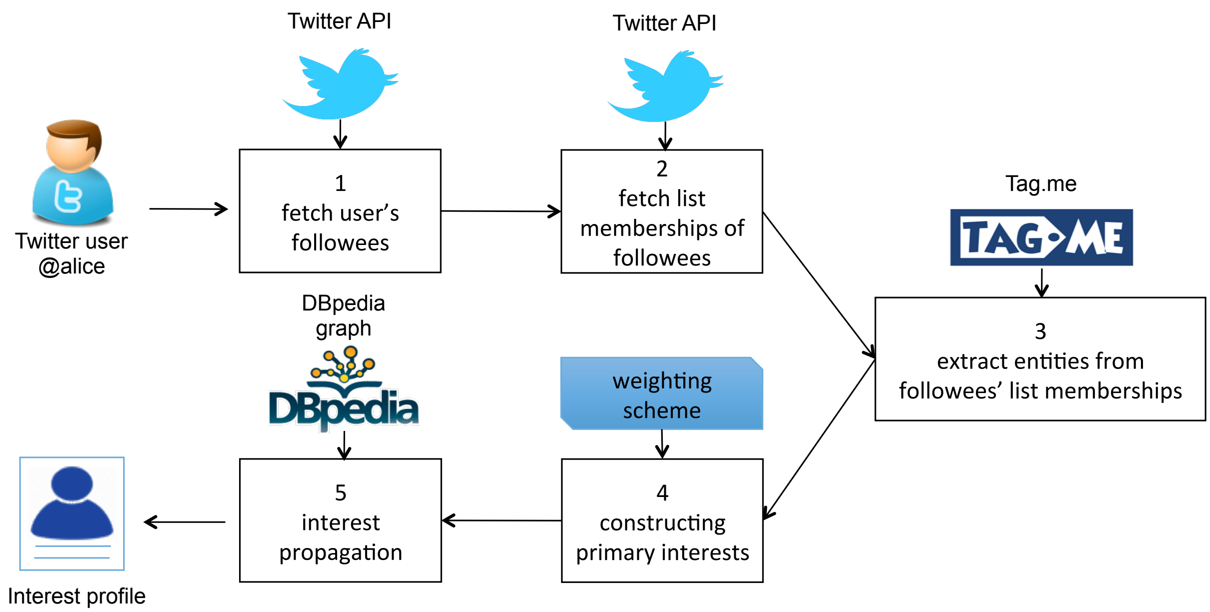



**Figure 2: Overview of user modeling strategy based on followees' list memberships.**

occurrences of an entity in a single followee's *list memberships*, but only counts the number of followees who have the entity in their profile. For example, the weight of an entity $c_j$ equals 5 if there are five followees of $u$ having the entity in their *list memberships*.

### 3.2 Interest Propagation Strategy

Based on the *primary interests* constructed in previous steps, background knowledge from DBpedia can be exploited to propagate user interests. For instance, we can assume that a user might be interested in `Apple_Inc.` if the user is interested in `Steve_Jobs` based on the corresponding categories of the entity `Steve_Jobs` from DBpedia.

Some *discounting strategies* can be used to discount the weights of propagated user interests based on the *primary interests* [14, 22]. We adopt the propagation method from [22] as the method showed overall better performance compared to the approach applying a spreading activation function from [14] in the previous study [25]. The proposed method [22] discounts a propagated category using the log scale of the numbers of sub-pages ($SP$) and sub-categories ($SC$, see Algorithm 4) of the category. The intuition behind this is that general categories, which have many sub-pages and sub-categories, should be discounted heavily.

$$CategoryDiscount = \frac{1}{\alpha} \times \frac{1}{\log(SP)} \times \frac{1}{\log(SC)} \tag{4}$$

Also, a propagated entity via a property is discounted based on the log scale of the number of occurrences of the property in the DBpedia graph ($P$, see Algorithm 5), i.e., if the property appears frequently in the graph, the entities extended through this property should be discounted heavily. In addition, $\alpha$ is a decay factor for the propagation from directly extracted entities to related categories or entities ($\alpha = 2$ as in the study [22]).

$$PropertyDiscount = \frac{1}{\alpha} \times \frac{1}{\log(P)} \tag{5}$$

**Extracting subset of DBpedia categories.** We consider leveraging all DBpedia categories of entities might be noisy since many Wikipedia categories are created for Wikipedia administration. Similar to the approach from [14], we extract a subset of all DBpedia categories which we use for our interest propagation. The subset consists of all inferred sub-categories of `dbc`[8]`:Main_topic_classifications`. However, different to [14] which requires the Wikipedia dump for extracting a hierarchical category graph, we connect directly to DBpedia to extract the subset of categories by using Algorithm 1. Therefore, it can be directly extracted via the DBpedia SPARQL Endpoint[9], and can be reproduced easily. In addition, we do not remove all administration categories (inferred sub-categories of `dbc:Wikipedia_administration`) as in [14] since we found that many useful categories are in the inferred sub-categories of the administration category as well as the main topic classification. This process results in 957,963 categories for our consideration while propagating user interests.

**Merging categories and entities with same title.** In DBpedia, many entities and categories have same title (name), e.g., `dbr`[10]`:Apple_Inc.` and `dbc:Apple_Inc.`. Considering these concepts separately as entities and categories might decrease the quality of propagated user interests or unnecessarily increase the size of user interest profiles. In this regard, we do not treat entities and

[8]The prefix `dbc` denotes http://dbpedia.org/resource/Category:
[9]http://dbpedia.org/sparql
[10]The prefix `dbr` denotes http://dbpedia.org/resource/

**Algorithm 1** GetSubsetOfDBpediaCategories

```
procedure getSubsetOfDBpediaCategories(topCategory)
    category_dictionary = {topCategory:0}            ▷ 0 denotes unprocessed
    while size(unprocessed categories in category_dictionary) > 0 do
        for category in unprocessed categories do
            if category not in category_dictionary then
                add category:0 to category_dictionary
    return keys of category_dictionary     ▷ return all inferred sub-categories
```

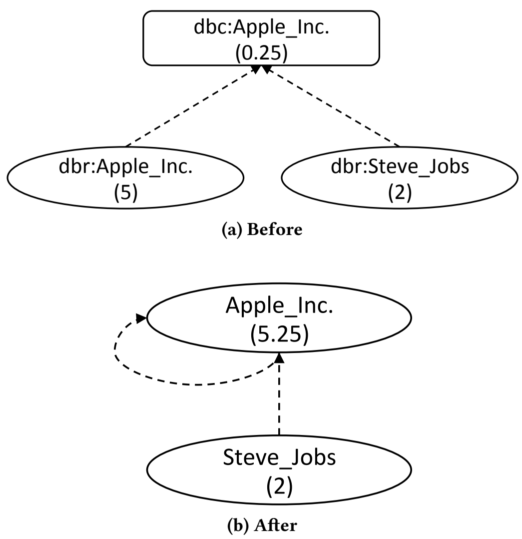



**Figure 3: Before and after merging categories and entities with the same title.**

categories differently in our propagation strategy, i.e, if there is a category which has same name with an entity that has been propagated, the category and entity will be merged into a single concept, and the weights will be accumulated. For example, in Figure 3(a), the propagated category `dbc:Apple_Inc.` has its own weight based on two entities `dbr:Apple_Inc.` and `dbr:Steve_Jobs` by considering categories and entities separately. On the other hand, Figure 3(b) shows that `Apple_Inc.` is considered as a single concept and its weight has been accumulated. In Section 5.2, we will show how these trimmed categories using Algorithm 1, and the strategy merging categories and entities with same title, positively affect the quality of inferred user interest profiles.

Finally, we apply Inverse Document Frequency (IDF) on the user interest profile $P_u$, and then normalize $P_u$ in order to make the sum of all concept weights equal to 1: $\sum_{c_i \in C} w(u, c_i) = 1$.

## 4 EXPERIMENTAL SETUP

In this section, we describe the evaluation methodology for evaluating constructed user interest profiles (Section 4.1), and the dataset used in our experiment (Section 4.2).

### 4.1 Evaluation Methodology

In the literature, user interest profiles have been evaluated in terms of recommendation performance for content-based recommendation systems by inputting different user interest profiles generated by different user modeling strategies [1, 2, 25, 31, 32]. In the same way, we evaluate different user interest profiles constructed based on different types of information (e.g, *bios* and *list memberships*) of followees in terms of a link (URL) recommendation system on Twitter. To this end, for a target user $u$, we construct the ground truth as the links shared via $u$'s tweets within the last two weeks. Afterwards, as our focus here is exploring different types of information of followees for inferring user interest profiles, we blind out all of $u$'s tweets, and use only different types information from followees of $u$ to build user interest profiles for $u$.

**A link (URL) profile** is constructed with the same representation model (i.e., Bag of Concepts) based on its content. For instance, DBpedia entities can be extracted based on the content of a link $l$, and the propagation strategy mentioned in Section 3.2 is then applied to those entities to build the link profile $P_l$. As our focus is not optimizing recommendation quality, we compare the quality of different user interest profiles with a lightweight recommendation algorithm when inputting different types of user interest profiles, similar to the one used in the previous studies [1, 2, 25].

**Recommendation Algorithm:** given a user interest profile $P_u$ and a set of candidate links $N = \left\{P_{l1}, ..., P_{ln}\right\}$, which are represented via profiles using the same vector representation, the recommendation algorithm ranks the candidate links according to their *cosine similarity* to $P_u$.

Therefore, the link recommender system provides the top-$N$ recommendations based on the cosine similarity scores between user and link profiles. Four evaluation metrics, namely *MRR* (Mean Reciprocal Rank), the success rate at rank $N$, recall at rank $N$, and precision at rank $N$ were used for evaluating the recommendation performance in the same way as previous studies [2, 4, 20, 22, 24]. We focus on $N = 10$ as our recommender system provides 10 link recommendations.

- **MRR** The *MRR* (Mean Reciprocal Rank) indicates at which rank the first link *relevant* to the user occurs (denoted by $rank_k$) on average.

$$MRR = \frac{1}{|U|} \sum_{k=1}^{|U|} \frac{1}{rank_k} \tag{6}$$

- **S@N** The Success at rank $N$ (S@$N$) stands for the mean probability that a relevant link occurs within the top-$N$ ranked.

$$S@N = \begin{cases} 1, & if\ a\ relevant\ link\ in \\ & retrieved\ links\ at\ N \\ 0, & otherwise \end{cases} \tag{7}$$

- **R@N** The Recall at rank $N$ (R@$N$) represents the mean probability that *relevant* links are successfully retrieved within the top-$N$ recommendations.

**Table 1: Dataset statistics**

| # of passive users | avg. # of considered followees | avg. # of list memberships of followees |
|---|---|---|
| 439 | 170 | 173 |

$$R@N = \frac{|\{relevant\ links\}| \cap |\{retrieved\ links\ at\ N\}|}{|\{relevant\ links\}|} \quad (8)$$

- **P@N** The Precision at rank $N$ (P@$N$) represents the mean probability that retrieved links within the top-$N$ recommendations are *relevant* to the user.

$$P@N = \frac{|\{relevant\ links\}| \cap |\{retrieved\ links\ at\ N\}|}{|\{retrieved\ links\}|} \quad (9)$$

We set the significance level of alpha as 5% for all statistical tests, and used the *bootstrapped paired t-test*[11] to test the significance.

### 4.2 Dataset

The Twitter dataset used in this study is from [21], which consists of 480 randomly chosen users on Twitter with their tweets and followees. We selected 439 users who have topical links (URLs which have at least four entities based on their content) in their tweets from last two weeks. All of the links shared by each user in the last two weeks of their timelines were used to build the set of candidate links for recommendations. On average, each user has 2,771 followees. As the rate limits of the Twitter API for retrieving followees and *list memberships* are 15 and 75 for a 15-minute window, we only consider up to 200 followees for each user, and crawled all *list memberships* of those followees for this study. The main details of our dataset are presented in Table 1. Finally, the dataset corresponds to 74,488 followees for 439 users with 170 followees on average, and the candidate set of links consists of 15,053 distinct links.

## 5 COMPARISON BETWEEN USING LIST MEMBERSHIPS AND BIOGRAPHIES OF FOLLOWEES FOR INFERRING USER INTERESTS

We use the recent approach which exploits *bios* of followees for inferring user interests [25] as a baseline to evaluate our user modeling strategies since the approach performs better than other approaches such as linking followee accounts to Wikipedia/DBpedia entities or leveraging the tweets of followees for inferring user interests.

[11] http://www.sussex.ac.uk/its/pdfs/SPSS_Bootstrapping_22.pdf

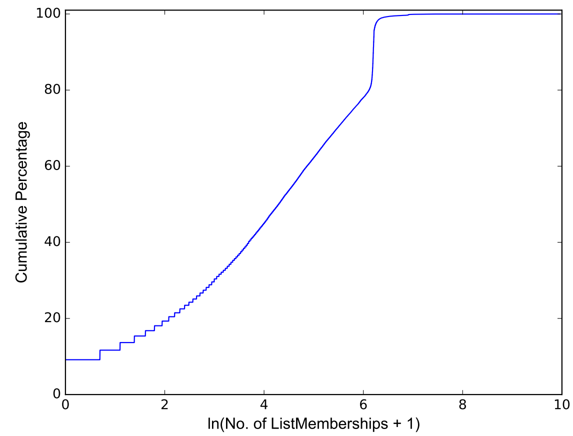



**Figure 4: Cumulative distribution of the number of list memberships of followees in the dataset.**

### 5.1 Quantitative analysis

We first look at how many list memberships a followee has been added into. The Cumulative Distribution Function (CDF) of the number of *list memberships* for 74,488 followees is shown in Figure 4. The figure shows that 90% of followees have less than 492 (ln(492+1)=6.2) *list memberships*. 6,871 (9.2%) out of 74,488 followees have no *list membership*, i.e., over 90% of followees have at least one *list membership*. On average, each followee belongs to 173 *list memberships*, which might be a useful information source of "descriptions" about a followee compared to the bio provided by him/her. For example, 3,047 entities can be extracted from the *list memberships* of followees on average when we consider up to 50 followees for each target user in our dataset. On the other hand, 23 entities can be extracted from the *bios* of followees on average. Given this quantified information from *list memberships* of followees, we move on to investigate whether it can be leveraged for building *qualified* user interest profiles in the context of link recommendations.

### 5.2 Qualitative analysis

Table 2 shows the link recommendation performance using three different user modeling strategies in terms of MRR, R@10, P@10, and S@10 respectively.

**Comparison between the baseline and our approach.** As we can see from the table, the user modeling strategy which exploits *list memberships* of followees using weighting scheme 1 ($UM(f\_listmemberships, WS1)$) performs better than $UM(f\_listmemberships, WS2)$ and the baseline method $UM(f\_bios)$. For example, when a passive user has less than 50 users, a significant improvement of $UM(f\_listmemberships, WS1)$ over $UM(f\_bios)$ in MRR (+17%, $p < 0.01$), P@10 (+12%, $p < 0.05$), and S@10 (+14%, $p < 0.05$) can be noticed. However, we can also observe that with the number of followees of a user increasing, the difference between using $UM(f\_listmemberships, WS1)$ and

**Table 2: Recommendation performance of different user modeling strategies in terms of four different evaluation metrics and numbers of followees. The best performing user modeling strategy is in bold. $^{**}$ denotes $p < 0.01$, and $^{*}$ denotes $p < 0.05$.**

| # of followees | Evaluation metric | UM(f_bios) [baseline] | UM(f_list memberships, WS1) | UM(f_list memberships, WS2) |
|---|---|---|---|---|
| 50 | MRR | 0.2243 | **0.2622** $^{**}$ | 0.2584 $^{*}$ |
| | R@10 | 0.0473 | **0.0532** | 0.0471 |
| | P@10 | 0.1226 | **0.1371** $^{*}$ | 0.1223 |
| | S@10 | 0.3690 | **0.4191** $^{*}$ | 0.4169 $^{*}$ |
| 100 | MRR | 0.258 | **0.2792** | 0.2613 |
| | R@10 | 0.0532 | **0.0584** | 0.0550 |
| | P@10 | 0.1428 | **0.1481** | 0.1337 |
| | S@10 | 0.4146 | **0.4579** $^{*}$ | 0.4442 |
| 150 | MRR | 0.2871 | **0.2995** | 0.2643 |
| | R@10 | 0.0579 | **0.0635** | 0.0609 |
| | P@10 | **0.1535** | 0.1508 | 0.1358 |
| | S@10 | 0.4579 | **0.4852** | 0.4738 |
| 200 | MRR | 0.2952 | **0.3065** | 0.2638 |
| | R@10 | 0.0627 | **0.0653** | 0.0575 |
| | P@10 | **0.1615** | 0.1526 | 0.1353 |
| | S@10 | 0.4715 | **0.4920** | 0.4784 |

$UM(f\_bios)$ becomes smaller. This shows that exploiting *list memberships* of followees can help with inferring user interest profiles in the case of a user having a small number of followees, which would be typical of "new" passive users.

**Comparison between two weighting schemes in our approach.** Table 2 also shows that the weighting scheme WS1 always outperforms WS2 in terms of four different evaluation metrics and different numbers of followees. The result indicates that WS1, which applies the normalized sum of occurrences of an entity from *list memberships* of followees, reflects the importance of the entity to passive users better when compared to the second weighting scheme which uses the number of followees having the entity in their *list memberships* (WS2).

**Effects of DBpedia refinement.** In Section 3.2, we introduced an interest propagation strategy by refining DBpedia categories and entities. Figure 5 shows the numbers of entities with/without refinement in terms of different numbers of followees. We found that the refinement of trimming DBpedia categories as well as merging categories and entities with the same name can compress the size of user interest profiles by around 9% compared to the user modeling strategy without the refinement, while remaining at a similar performance level in the context of link recommendations.

Another observation we noticed is that the recommendation results using *biographies* and *list memberships* might complement each other. For different users, we found that using *biographies*

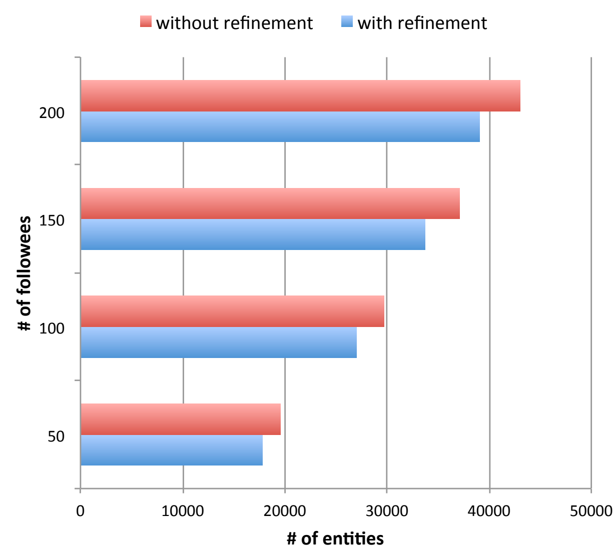



**Figure 5: Number of entities in terms of different number of followees of a user using WS1 with/without refinement.**

**Table 3: Recommendation performance of combining two views (from bios and list memberships) of followees compared to the baseline in terms of four different evaluation metrics and numbers of followees. The best performing user modeling strategy is in bold. ** denotes $p < 0.01$, and * denotes $p < 0.05$.**

| # of followees | Evaluation metric | UM(f_bios) [baseline] | Combined |
|---|---|---|---|
| 50 | MRR | 0.2243 | **0.2777** ** |
| | R@10 | 0.0473 | **0.0475** |
| | P@10 | 0.1226 | **0.1396** ** |
| | S@10 | 0.3690 | **0.4305** ** |
| 100 | MRR | 0.258 | **0.2946** ** |
| | R@10 | 0.0532 | **0.0584** * |
| | P@10 | 0.1428 | **0.1615** ** |
| | S@10 | 0.4146 | **0.4784** ** |
| 150 | MRR | 0.2871 | **0.3303** ** |
| | R@10 | 0.0579 | **0.0639** * |
| | P@10 | 0.1535 | **0.1745** ** |
| | S@10 | 0.4579 | **0.5194** ** |
| 200 | MRR | 0.2952 | **0.3397** ** |
| | R@10 | 0.0627 | **0.0654** |
| | P@10 | 0.1615 | **0.1779** ** |
| | S@10 | 0.4715 | **0.5125** * |

provides better performance while using *list memberships* does not and vice versa. To test the hypothesis whether combining two different views about followees can improve the quality of user modeling or not, we use an approach used in the literature in the next section.

## 6 COMBINING TWO VIEWS OF FOLLOWEES

As we mentioned in Section 1, the *bio* of a followee $f$ can be seen as a *self-description* of himself / herself, while the *list memberships* of $f$ can be seen as *others-descriptions* about *f*. In this section, we investigate whether combining these two different views of followees can complement each other in terms of the recommendation performance.

To this end, we apply a simple method used in [30], which is based on the principle of *polyrepresentation* [13]. The approach [30] combined different views of a user for predicting user interests in the context of a search engine. The final rank of an item is determined by the average rank position of each rank based on $UM(f\_bios)$ and $UM(f\_listmemberships, WS1)$. For example, if an item $i$ is ranked in x-th and y-th position based on $UM(f\_bios)$ and $UM(f\_listmemberships, WS1)$, the combined score for the item $i$ is $1/(x + y)$. The higher the value is, the higher the item will be ranked. We also evaluated an alternative approach for combining the two views which puts them into a single vector for building user interest profiles. However, the simple approach used in [30] provides better performance than the alternative. Therefore, we report the results based on [30] in this section.

The recommendation performance of user modeling strategy combining two different views (*self-descriptions* and *others-descriptions*) of followees compared to the baseline user modeling strategy using bios (*self-descriptions* only) of followees is displayed in Table 3. As we can see from the table, combining two different views with a simple approach clearly outperforms the baseline method significantly in terms of four different evaluation metrics. Also, while using *list memberships* of followees only has a significant difference compared to the baseline when the number of followees is small (i.e., # of followees = 50, 100, see Table 2), the combined approach has a higher significant difference ($p < 0.01$) compared to the baseline method even when the number of followees becomes larger (i.e., # of followees = 100, 150, 200, see Table 3).

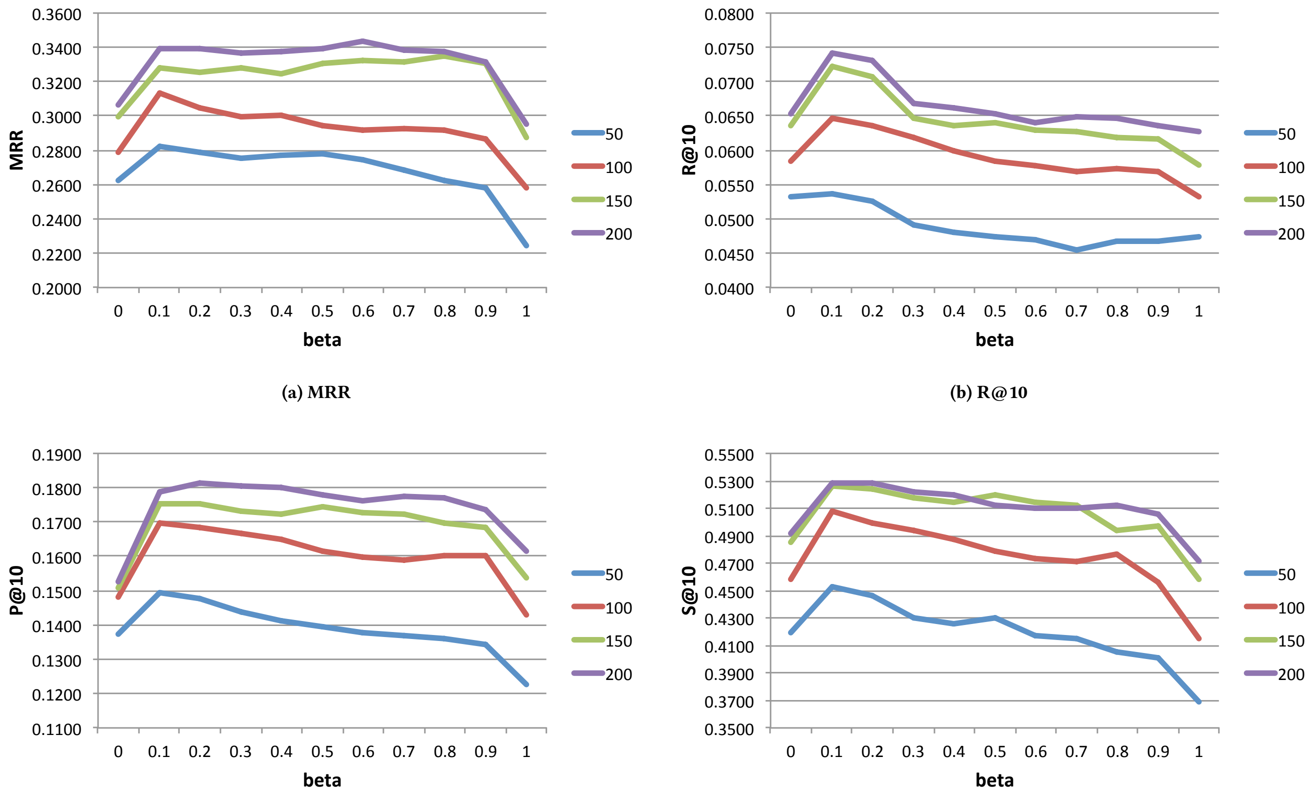
The aforementioned combination of the two views considers the importance of each view equally [30]. To better understand which view of followees has higher importance in different situations, we change the combined score as $1/(\beta \times x + (1 - \beta) \times y)$, where $\beta$ controls the importance of the first view, i.e., *bios* (*self-descriptions*) of followees. As one might expect, $\beta = 0$ denotes that we only consider *list memberships* (*other-descriptions*) of followees, while $\beta = 1$ denotes that we only consider *bios*(*self-descriptions*) of followees. $\beta = 0.5$ denotes that we treat two different views of followees equally as we already discussed earlier in this section.

Figure 6 shows the link recommendation performance in terms of four evaluation metrics by setting $\beta$ between 0 and 1 in steps of 0.1. As depicted in Figure 6, the recommendation performance is better with smaller values of $\beta$ for combining the two different views (i.e., *self-descriptions* and *others-descriptions*) of followees for inferring user interest profiles in terms of R@10, P@10 and S@10. The best performance is achieved with $\beta = 0.1$, and the performance starts decreasing with increasing $\beta$. This denotes *others-descriptions* of followees plays more important role for combining the two views. Similar results can be observed in terms of MRR with a small number of followees, i.e., # of followees = 50 or 100. However, as we can see from Figure 6 (a) that, with a big number of followees, i.e., # of followees = 150 or 200, the difference with different $\beta$ values are tending towards being stable in terms of MRR.

Thus we conclude that *bios* (*self-descriptions*) and *list memberships* (*others-descriptions*) can complement each other and improve the quality of user modeling in terms of link recommendations. Also, *list memberships* plays a more important role for combining the two different views especially in the case of a small number of followees being discovering available.

## 7 CONCLUSIONS

In this paper, we were interested in whether leveraging *list memberships* of followees can provide quantitative and qualitative information for inferring user interests for *passive users*, which has not been studied before. In addition, we further investigated whether the two different views, *biographies* (*self-descriptions*) and *list memberships*




(a) MRR

(b) R@10

(c) P@10

(d) S@10

**Figure 6: The quality of user modeling with different $\beta$ values for combining two different views (*self-descriptions* and *others-descriptions*) of followees in terms of link recommendations on Twitter.**

(*others-descriptions*) of followees, can complement each other to improve the quality of inferred user interest profiles. A series of offline experiments were performed to evaluate the inferred user interest profiles built by different user modeling strategies in terms of a link (URL) recommender system on Twitter using four different evaluation metrics. The study results indicate that: (1) leveraging *list memberships* of followees performs better than exploiting *biographies* especially in the case of a user having a small number of followees, (2) combining the two different views of followees can improve the quality of user modeling significantly compared to the baseline method which exploits *biographies* of followees only, and the *list memberships* of followees play a more important role in the combination. As a further step, we plan to study whether combining other views of followees, such as their generated content, can also have a synergetic effect on user modeling to improve the inferred user interest profiles for passive users.

## 8 ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight Centre for Data Analytics).

## REFERENCES

[1] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. 2011. Analyzing Temporal Dynamics in Twitter Profiles for Personalized Recommendations in the Social Web. In *Proceedings of the 3rd International Web Science Conference*. ACM, 2.

[2] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. 2011. Analyzing User Modeling on Twitter for Personalized News Recommendations. In *User Modeling, Adaption and Personalization*. Springer, 1–12.

[3] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. 2011. Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. In *The Semantic Web: Research and Applications*. Springer, 375–389.

[4] Fabian Abel, Claudia Hauff, Geert-Jan Houben, and Ke Tao. 2012. Leveraging User Modeling on the Social Web with Linked Data. In *Web Engineering SE - 31*. Springer, 378–385.

[5] Amr Ahmed, Yucheng Low, Mohamed Aly, Vanja Josifovski, and Alexander J Smola. 2011. Scalable distributed inference of dynamic user interests for behavioral targeting. In *Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining*. ACM, 114–122.

[6] Christoph Besel, Jörg Schlötterer, and Michael Granitzer. 2016. Inferring Semantic Interest Profiles from Twitter Followees: Does Twitter Know Better Than Your Friends?. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing (SAC '16)*. ACM, New York, NY, USA, 1152–1157. https://doi.org/10.1145/2851613.2851819

[7] Christoph Besel, Jörg Schlötterer, and Michael Granitzer. 2016. On the Quality of Semantic Interest Profiles for Onine Social Network Consumers. *SIGAPP Appl. Comput. Rev.* 16, 3 (nov 2016), 5–14. https://doi.org/10.1145/3015297.3015298

[8] Ceren Budak, Anitha Kannan, Rakesh Agrawal, and Jan Pedersen. 2014. *Inferring User Interests from Microblogs.* Technical Report. Microsoft.

[9] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. 2010. Short and Tweet: Experiments on Recommending Content from Information Streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

ACM, 1185–1194.
[10] Stefano Faralli, Giovanni Stilo, and Paola Velardi. 2015. Recommendation of Microblog Users based on Hierarchical Interest Profiles. *Social Network Analysis and Mining* 5, 1 (2015), 1–23.
[11] John Hannon, Kevin McCarthy, Michael P O'Mahony, and Barry Smyth. 2012. A multi-faceted user model for twitter. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 303–309.
[12] Morgan Harvey, Fabio Crestani, and Mark J Carman. 2013. Building User Profiles from Topic Models for Personalised Search. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2309–2314. https://doi.org/10.1145/2505515.2505642
[13] Peter Ingwersen. 1994. Polyrepresentation of information needs and semantic entities elements of a cognitive theory for information retrieval interaction. In *SIGIR'94*. Springer, 101–110.
[14] Pavan Kapanipathi, Prateek Jain, Chitra Venkataramani, and Amit Sheth. 2014. User Interests Identification on Twitter Using a Hierarchical Knowledge Base. In *The Semantic Web: Trends and Challenges*. Springer, 99–113. https://doi.org/10.1007/978-3-319-07443-6_8
[15] Dongwoo Kim, Yohan Jo, Il-Chul Moon, and Alice Oh. 2010. Analysis of twitter lists as a potential source for discovering latent characteristics of users. In *ACM CHI Workshop on Microblogging*. Citeseer, 4.
[16] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, and Sören Auer. 2013. Dbpedia-a Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal* (2013).
[17] Matthew Michelson and Sofus A Macskassy. 2010. Discovering Users' Topics of Interest on Twitter: A First Look. In *Proceedings of the 4th Workshop on Analytics for Noisy Unstructured Text Data*. ACM, 73–80.
[18] Alan Mislove, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel. 2010. You are Who You Know: Inferring User Profiles in Online Social Networks. In *Proceedings of the third ACM International Conference on Web Search and Data Mining*. ACM, 251–260.
[19] Fedelucio Narducci, Cataldo Musto, Giovanni Semeraro, Pasquale Lops, and Marco Gemmis. 2013. Leveraging Encyclopedic Knowledge for Transparent and Serendipitous User Profiles. In *User Modeling, Adaptation, and Personalization: 21th International Conference*. Springer Berlin Heidelberg, Berlin, Heidelberg, 350–352.
[20] Fabrizio Orlandi, John Breslin, and Alexandre Passant. 2012. Aggregated, Interoperable and Multi-domain User Profiles for the Social Web. In *Proceedings of the 8th International Conference on Semantic Systems*. ACM, 41–48. https://doi.org/10.1145/2362499.2362506
[21] Guangyuan Piao and John G. Breslin. 2016. Analyzing Aggregated Semantics-enabled User Modeling on Google+ and Twitter for Personalized Link Recommendations. In *User Modeling, Adaptation, and Personalization*. ACM, 105–109.
[22] Guangyuan Piao and John G Breslin. 2016. Exploring Dynamics and Semantics of User Interests for User Modeling on Twitter for Link Recommendations. In *12th International Conference on Semantic Systems*. ACM, 81–88.
[23] Guangyuan Piao and John G Breslin. 2016. Interest Representation, Enrichment, Dynamics, and Propagation: A Study of the Synergetic Effect of Different User Modeling Dimensions for Personalized Recommendations on Twitter. In *Knowledge Engineering and Knowledge Management: 20th International Conference*. Springer International Publishing, 496–510.
[24] Guangyuan Piao and John G. Breslin. 2016. User Modeling on Twitter with WordNet Synsets and DBpedia Concepts for Personalized Recommendations. In *The 25th ACM International Conference on Information and Knowledge Management*. ACM, 2057–2060.
[25] Guangyuan Piao and John G. Breslin. 2017. Inferring User Interests for Passive Users on Twitter by Leveraging Followee Biographies. In *39th European Conference on Information Retrieval*. Springer.
[26] Zhonghua Qu and Yang Liu. 2011. Interactive group suggesting for Twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 519–523.
[27] Amit Sheth and Pavan Kapanipathi. 2016. Semantic Filtering for Social Data. *IEEE Internet Computing* 20, 4 (2016), 74–78.
[28] Patrick Siehndel and Ricardo Kawase. 2012. TwikiMe!: user profiles that make sense. In *Proceedings of the 2012th International Conference on Posters & Demonstrations Track-Volume 914*. CEUR-WS. org, 61–64.
[29] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short Text Classification in Twitter to Improve Information Filtering. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 841–842.
[30] Ryen W White, Peter Bailey, and Liwei Chen. 2009. Predicting User Interests from Contextual Information. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*. ACM, New York, NY, USA, 363–370. https://doi.org/10.1145/1571941.1572005
[31] Fattane Zarrinkalam. 2015. Semantics-Enabled User Interest Mining. In *The Semantic Web. Latest Advances and New Domains SE - 54*, Fabien Gandon, Marta Sabou, Harald Sack, Claudia D'Amato, Philippe Cudré-Mauroux, and Antoine Zimmermann (Eds.). Lecture Notes in Computer Science, Vol. 9088. Springer International Publishing, 817–828. https://doi.org/10.1007/978-3-319-18818-8_54
[32] Fattane Zarrinkalam and Mohsen Kahani. 2015. Semantics-enabled User Interest Detection from Twitter. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 469–476.