***This paper has been accepted** for presentation at the 12th International Conference on Semantic Systems (SEMANTiCS2016) in Leipzig, Germany.

Exploring Dynamics and Semantics of User Interests for User Modeling on Twitter for Link Recommendations

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

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

ABSTRACT

User modeling for individual users on the Social Web plays an important role and is a fundamental step for personalization as well as recommendations. Recent studies have proposed different user modeling strategies considering various dimensions such as *temporal dynamics* and *semantics* of user interests. Although previous work proposed different user modeling strategies considering the temporal dynamics of user interests, there is a lack of comparative studies on those methods and therefore the comparative performance over each other is unknown. In terms of semantics of user interests, background knowledge from DBpedia has been explored to enrich user interest profiles so as to reveal more information about users. However, it is still unclear to what extent different types of information from DBpedia contribute to the enrichment of user interest profiles.

In this paper, we propose user modeling strategies which use Concept Frequency - Inverse Document Frequency (CF-IDF) as a weighting scheme and incorporate either or both of the *dynamics* and *semantics* of user interests. To this end, we first provide a comparative study on different user modeling strategies considering the dynamics of user interests in previous literature to present their comparative performance. In addition, we investigate different types of information (i.e., *categories*, *classes* and connected *entities* via various properties) for entities from DBpedia and the combination of them for extending user interest profiles. Finally, we build our user modeling strategies incorporating either or both of the best-performing methods in each dimension. Results show that our strategies outperform two baseline strategies significantly in the context of link recommendations on Twitter.

1. INTRODUCTION

Personalization on the Web started by analyzing Web documents that users visit in order to generate users' interests [9, 24]. Recently, with the increasing adoption of Online Social Networks (OSNs) such as Twitter[1], service providers need to analyze user activities on their platforms to provide personalized services for users. An important aspect of personalization is the creation of a high quality user profile that provides an accurate representation of the user interests via user modeling [24]. *Bag of Concepts* approach, e.g, using DBpedia [16] entities to represent user interests, has been widely used to represent user interests on the Social Web. Throughout the rest of paper, by a *concept* we mean an *entity*, *category* or *class* from a Knowledge Base (KB) for representing user interests.

Example tweet: *My Top 3 #lastfm Artists: Eagles of Death Metal, The Black Keys & The Wombats #mm bit.ly/dcwe7t.* Given this tweet posted by a user, we know that the user is interested in entities such as `dbpedia`[2]`:The_Wombats` and `dbpedia:The_Black_Keys`. We term the entities that can be directly extracted from a user's tweets as ***primitive interests***.

On top of the concept-based representation of user interests, researchers explored various dimensions of user modeling to better understand user interests and improve the quality of user modeling. *Dynamics* and *semantics* of user interests as two major design dimensions of user modeling on the Social Web have been studied in the past few years.

Dynamics of user interests. With an underlying assumption that the interests of users might change over time, several methods have been proposed [2, 20] to incorporate the temporal dynamics of user interests. Previous studies showed that considering the time decay of user interests improves the quality of user modeling in the context of personalized recommendations. However, there was no study providing a comparison of different methods proposed in the literature.

Semantics of user interests. To better understand semantics of User-Generated Content (UGC) on the Social Web such as Twitter, Linked Open Data (LOD) [12], especially DBpedia [16] which is a 1st-class citizen of the LOD cloud, has been exploited since it provides a great variety of information that can support various applications [7], including semantic enrichment of tweets [15] and user modeling [5]. One of the benefits of using concepts from a KB such as DBpedia for representing user interests is that user interests can be further extended by leveraging background knowledge with respect to the concepts from the Knowledge Base. For instance, in the example tweet, we can further infer that the user is interested in `dbpe-`

WOODSTOCK '97 El Paso, Texas USA

DOI: 10.475/123_4

[1] http://twitter.com

[2] The prefix `dbpedia` denotes http://dbpedia.org/resource/

dia:Indie_rock as both dbpedia:The_Wombats and dbpedia:The_Black_Keys are pointing to dbpedia:Indie_rock via the property dbpedia-owl[3]:genre. We term the concepts propagated from *primitive interests* as ***propagated interests***. To better exploit background knowledge from DBpedia, different types of information (e.g., *categories*, *classes* and *connected entities via different properties*) should be considered all together for propagating user interest profiles.

The contributions of this work are summarized as follows.

- We use Concept Frequency - Inverse Document Frequency (CF-IDF) as a weighting scheme for user modeling in our study and show the improvement over the CF weighting scheme (see Section 5.1).
- We provide a comparative study on different methods proposed in the literature for incorporating the dynamics of user interests on the Social Web (see Section 5.2). To our knowledge, this is the first comparative study on those methods.
- We investigate different types of information from DBpedia, i.e., categories, classes, connected entities via different properties and the combination of them, for extending user interests (see Section 5.3).
- Finally, we evaluate our user modeling strategies, which incorporate either or both of the dynamics and semantics of user interests, by comparing two baseline strategies in the literature in the context of link recommendations on Twitter (see Section 5.4).

The organization of the rest of the paper is as follows. Section 2 gives some related work, and Section 3 describes our user modeling framework. In section 4, we present the experiment setup for our study. Section 5 investigates the weighting scheme of user interest profiles, and two dimensions of user modeling in detail. We evaluate our proposed user modeling strategies in Section 5, and conclude in Section 6 with some future work.

2. RELATED WORK

To represent user interests, previous work either used *Bag of Words* [19], *Topic Model* [11] or *Bag of Concepts* [3, 4, 20] approach. *Bag of Words* and *Topic Model* focus on words, which cannot provide semantic information and relationships among words. In addition, the assumption behind *Topic Model* is that a single document contains rich information. This is not the case in OSNs such as Twitter, and as a result, decreases the quality of user modeling [14, 18, 25]. Abel et al. [3] showed that *entity-based* user profiles outperform other approaches such as *hashtag-* and *topic-based* user profiles on Twitter in the context of news recommendations. In contrast to word-based approaches, *Bag of Concepts* approach, e.g, using DBpedia entities for representing user interests, can exploit background knowledge linked to the concepts to enrich user interest profiles [20, 22]. Therefore, we use the *Bag of Concepts* approach in this work and focus on user modeling strategies on top of this approach. In the rest of this section, we give related work on two dimensions of user modeling - temporal dynamics and semantics of user interests on the Social Web.

[3]The prefix dbpedia-owl denotes http://dbpedia.org/ontology/

2.1 Temporal dynamics of user interests

Based on the hypothesis that the interests of users change over time, previous work considered the temporal dynamics of interests for user modeling in OSNs [2,3,6,8,20]. In [3], the authors evaluated *short-term* and *long-term* user profiles in the context of news recommendations on Twitter. Short-term user profiles extract user interests within a short-term period (e.g., the last two weeks) while the long-term user profiles extract user interests from entire historical UGC of users. Instead of extracting user interest profiles within a certain period, researchers have proposed using an interest decay function for the interests of users. In this case, the weights of user interests were discounted by time, i.e., the interests appearing a long time ago would decay heavily. For evaluation, every work proposed its own evaluation method for its measure and none of those studies compared their proposed methods to others. For example, some of them evaluated their approaches based on a user study [20] while others evaluated them in terms of personalized recommendation systems in comparison to the method without considering the dynamics of user interests [2,3,6]. In this work, we provide a comparative evaluation over those methods in the context of personalized link recommender systems to examine the pros and cons of each measure (see Section 5.2).

2.2 Leveraging background knowledge for user modeling

On top of the *entity-based* user profiles, researchers have proposed using rich semantic information from a KB to extend user interests. For example, Abel et al. [5] proposed exploiting DBpedia to extend user profiles with respect to point of interests (POI). Their experiment showed that the extended POI profiles outperform the original user profiles without any extension in the context of POI recommendations. Different from focusing on user interests in a specific domain, our work focuses on user interests extracted from Twitter which are not limited to a specific domain. Orlandi et al. [20] proposed *category-based* user profiles based on the category information for entities from DBpedia. In addition to a straightforward extension giving equal weight to each extended category with respect to an entity, they proposed a discounting strategy for the extended categories. On top of that, Piao et al. [22] proposed a mixed approach that combines the entity- and category-based user profiles with a discounting strategy and proved that the mixed approach performs better than either the entity- or category-based approach. Our work differs from previous work in several aspects. First, we adopt CF-IDF weighting scheme instead of CF one. Secondly, instead of using category information for extending user interests, we consider three different types of information (i.e., *categories*, *classes* and connected *entities* via various properties) as well as the combination of them.

3. CONTENT-BASED USER MODELING

In this work, we use DBpedia concepts for representing the interests of users. The generic model for profiles representing users is specified in Definition 1.

Definition 1. The interest profile of a user $u \in U$ is a set of weighted DBpedia concepts. The weight with respect to the given user u for a concept $c \in C$ is computed by a certain function $w(u, c)$.

$$P_u = \{(c, w(u,c)) \mid c \in C, u \in U\} \quad (1)$$

Here, C and U denote the set of concepts in DBpedia and users respectively.

The process of generating user interest profiles on Twitter is displayed in Figure 1. It has three major steps:

(1) Primitive interests extraction. For a given user, we extract all DBpedia entities (*primitive interests*) within UGC of the user using the `Aylien API`[4]. For instance, two entities `dbpedia:Google_X` and `dbpedia:Cancer` can be retrieved from the phrase: "Google[x] Reveals Nano Pill To Seek Out Cancerous Cells Detecting cancer could be as easy as popping a pill in the near future". Concept Frequency (CF) is applied to denote the importance of a concept for a user. In addition, it might adhere to strategies for incorporating the temporal dynamics of user interests. Stop entities like `RT_(Network)` for `@RT` in tweets are removed while extracting primitive interests for users.

(2) Interest propagation. In this step, our user modeling framework might apply propagation strategies for primitive interests with background knowledge from DBpedia. The output here is a user interest profile consisting of *primitive interests* as well as *propagated interests*.

(3) Weighting and normalization. Finally, the user modeling framework applies Inverse Document Frequency (IDF) on the user interest profile, and further normalizes the profile so that the sum of all weights in the profile is equal to 1: $\sum_{c_i \in E} w(u, c_i) = 1$.

The strategies implemented in our user modeling framework for incorporating the temporal dynamics of user interests, and for extending user interest profiles with background knowledge, are investigated in detail in Section 5.

4. EXPERIMENT SETUP

In this section, we describe the dataset for our study (Section 4.1) and the evaluation methodology (Section 4.2) for the experiment.

4.1 Dataset

We used the Twitter dataset from [22], which includes all tweets published by 480 *active* users on Twitter (a user is *active* if the user published at least 100 posts [13, 17, 22]). The main details of the dataset are presented in Table 1. We further selected users who shared at least one link (URL) in their tweets during the last two weeks. We only consider links having at least four topics (concepts) to filter out non-topical links (e.g., links sharing current location via Swarm[5]). 322 out of 480 users met the criteria who published 247,676 tweets in total.

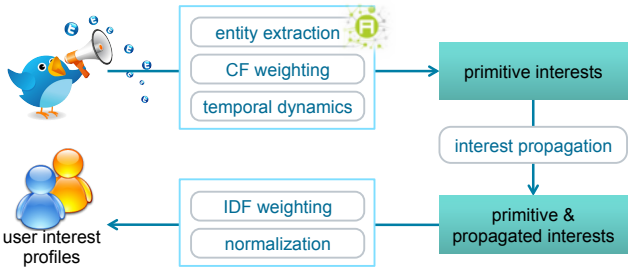


Figure 1: The process of generating user interest profiles on Twitter

Table 1: Dataset statistics

# of users	480
total # of tweets	348,554
average time span of tweets per user (days)	471
average # of tweets per user	726
average # of tweets per user per day	7.2

4.2 Evaluation methodology

Our main goal here is to analyze and compare the different user modeling strategies in the context of link recommendations. We do not aim to optimize the recommendation quality, but are interested in comparing the quality achieved by the same recommendation algorithm when inputting user profiles based on different user modeling strategies. In the same way from literature, we adopt a lightweight content-based algorithm as the recommendation algorithm that recommends links according to their *cosine* similarity with a given user profile [2].

Definition 2. Recommendation Algorithm: given a user profile P_u and a set of candidate links $\mathrm{N} = \{P_{i1}, ..., P_{in}\}$, which are represented via profiles using the same vector representation, the recommendation algorithm ranks the candidate items according to their cosine similarity to the user profile.

The ground truth of links, which we consider as *relevant* for a specific user, was given by links shared via the user's tweets within the last two weeks. We used the ground truth links from 322 users, as well as the links shared by other users but not shared by 322 users in the dataset, for constructing candidate links. In total, the ground truth of links consists of 3,959 links and the candidate set of links consists of 15,440 distinct links. The rest of tweets before the recommendation time were all used for constructing user profiles.

Given the ground truth and the candidate set of links, we applied different user modeling strategies together with the recommendation algorithm (see Definition 2) to provide personalized link recommendations. The quality of the top-N recommendations was measured via the following metrics, which have been used in previous studies [3, 5, 20].

- **MRR** The *MRR* (Mean Reciprocal Rank) indicates at which rank the first item *relevant* to the user occurs on average.
- **S@N** The Success at rank N (S@N) stands for the mean probability that a relevant item occurs within the top-N of the ranking.
- **R@N** The Recall at rank N (R@N) represents the mean probability that *relevant* items are successfully retrieved within the top-N recommendations.
- **P@N** The Precision at rank N (R@N) represents the mean probability that retrieved items within the top-N recommendations are *relevant* to the user.

[4]http://aylien.com

[5]https://www.swarmapp.com

We focused on $N = 10$ as our recommendation system will list 10 link recommendations to a user. We used the *bootstrapped paired t-test*[6] (which is an alternative to the paired t-test when the assumption of normality of the method is in doubt) for testing the significance where the significance level was set to 0.05 unless otherwise noted.

5. STUDY OF USER MODELING STRATEGIES

In this section, we first present CF-IDF weighting scheme and show its effectiveness on the quality of user modeling in the context of link recommendations (Section 5.1). Next, we provide a comparative study on different methods in the literature for considering the temporal dynamics in user modeling on the Social Web (Section 5.2). We then investigate various extension strategies using different types of information from DBpedia for user modeling in Section 5.3. Finally, we present the performance of link recommendations using our user modeling strategies compared to using two baseline methods (Section 5.4).

5.1 CF-IDF weighting scheme

The weighting scheme $w(u, c)$ measures the importance of a concept with respect to a user. Previous studies have been applied Concept Frequency (CF) as the weighting scheme $w_{CF}(u, c)$ for concept-based user profiles [2, 20]. The weight of a concept (interest) is determined by the number of OSN activities in which user u refers to the concept c. For instance, in a Twitter profile of user u, $w(u,$ `dbpedia:IPad`$) = 7$ means that u published seven Twitter messages that mention the entity `dbpedia:IPad`. In contrast, we make use of CF and Inverse Document Frequency (IDF) for our weighting scheme $w_{CF\text{-}IDF}(u, c)$, which was proposed and evaluated in the context of news recommender systems based on a user study [10]. Similar to the TF-IDF weighting scheme used in *word-based* user modeling approaches [1], the rationale behind CF-IDF is that concepts appearing in many users' interest profiles can be discounted while concepts appearing in a specific user's profile can obtain a higher weight. More formally, it is defined as follows.

- $w_{CF}(u, c) = the\ frequency\ of\ c\ in\ a\ user's\ tweets$,
- $w_{CF\text{-}IDF}(u, c) = \underbrace{w_{CF}(u, c)}_{CF} \times \underbrace{\log \frac{M}{m_c}}_{IDF}$

where M is the total number of users and m_c is the number of users interested in a concept c.

5.1.1 Results

As there was no comparison of CF and CF-IDF weighting schemes for user modeling on Twitter, we evaluated our choice of the weighting scheme in the context of link recommendations on Twitter. Figure 2 illustrates the recommendation performance of using CF and CF-IDF weighting schemes. As we can see from the figure, the weighting scheme $w_{CF\text{-}IDF}(u, c)$ clearly outperforms the $w_{CF}(u, c)$ in terms of all metrics and improves the recommendation performance significantly. Hence, we continue our experiment with $w_{CF\text{-}IDF}(u, c)$ as the weighting scheme for user interest profiles in the following experiment.

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

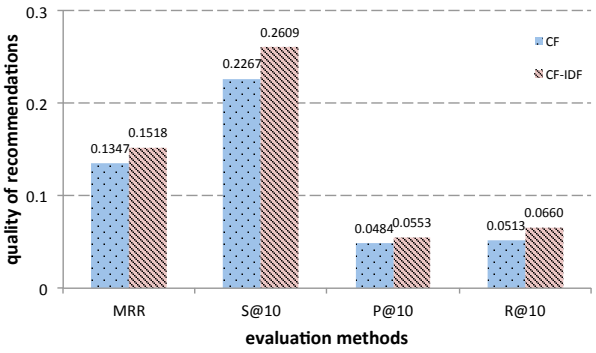


Figure 2: The quality of recommendations using CF and CF-IDF as the weighting schemes for user modeling

5.2 Comparison of different methods considering the dynamics of user interests

To study the effectiveness of different methods for incorporating dynamics of user interests, we implemented those methods mentioned in Section 2. In the following, `Long-term` denotes interest profiles that are generated based on all of the historical UGC of users while `Short-term` indicates interest profiles that are generated based on the last two weeks of users' UGC before the recommendation time. In the rest of this section, we describe each method in the literature in detail and evaluate them in the context of link recommendations on Twitter.

Interest decay functions. Based on `Long-term` user profiles, researchers have proposed different interest decay functions with the assumption that interests appearing recently should be given a higher weight than that appearing a long time ago. In the following, we use `Long-term(X)` to denote user profiles with a specific interest decay function `X`.

`Long-term(Orlandi)`: Orlandi et al. [20] proposed an exponential decay function for ranking user interests as follows:

$$x(t) = e^{-t/\beta} \tag{2}$$

In addition, they defined an initial time window (7 days) where the interests are not discounted by the decay function. We set the value of $\beta = 360 days$ in our experiment as in [20]. From a practical point of view, the interest decay function indicates that an interest value is discounted to 37% of its initial value (which is one by default) after 360 days.

`Long-term(Ahmed)`: Ahmed et al. [6] proposed getting the expected weight in terms of an interest k for user i at time t by combining three levels of abstractions using a weighted sum as below:

$$w_{ik}^{t} = \mu_{week} w_{ik}^{t,week} + \mu_{month} w_{ik}^{t,month} + \mu_{all} w_{ik}^{t,all} \tag{3}$$

where $\mu_{week} = \mu$, $\mu_{month} = \mu^2$ and $\mu_{all} = \mu^3$ where $\mu \in [0, 1]$. We set μ as e^{-1} in the same way from [6] for our experiment. As this method was proposed and evaluated in terms advertisement recommendations on web portals (i.e.,

Yahoo![7]), we modify μ_{week} and μ_{month} to μ_{2week} and μ_{2month} respectively to enable the method to be adapted to link recommendations on Twitter. The underlying assumption of the modification is that user interests decay slowly on Twitter as proved in based on a user study [20]. We use `Long-term(Ahmed`α`)` to denote the modified version of `Long-term(Ahmed)`. This interest decay function combines three levels of abstractions where the decay of user interests in each abstraction is μ times the previous abstraction. In contrast, user interests in Equation 2 (`Long-term(Orlandi)`) decay smoothly over time.

`Long-term(Abel)`: Abel et al. [2] proposed a time-sensitive interest decay function, which dampens the occurrence frequency of an entity e according to the temporal distance between the entity occurrence time and the given timestamp.

$$w(e, time, T_{tweets,u}) = \sum_{t \in T_{tweets,u,e}} (1 - \frac{|time - time(t)|}{max_{time} - min_{time}})^d \quad (4)$$

where $T_{tweets,u,e}$ denotes the set of tweets that have been published by a user u and refer to an entity e. *time(t)* returns the timestamp of a given tweet t and max_{time} and min_{time} denote the highest (youngest) and lowest (oldest) timestamp of a tweet in $T_{tweets,u,e}$. The parameter d is used to adjust the influence of the temporal distance. We set the parameter $d = 4$ as in [2]. As we can see from Equation 4, this approach not only considers how old an entity e is from the recommendation time but also incorporates the time span of the entity in the user's historical UGC.

5.2.1 Results

The results of the link recommendation experiment using different interest decay functions are summarized in Figure 3. In line with the result from [3], `Short-term` profiles do not outperform `Long-term` profiles. In terms of `Long-term(X)` user interest profiles, `Long-term(Ahmed)`, `Long-term(Ahmed`α`)` as well as `Long-term(Orlandi)` have comparative performance in terms of all evaluation metrics and perform significantly better than the user profiles without considering any decay of user interests (`Long-term`). `Long-term(Abel)` has slightly better performance in comparison to `Long-term` but the difference is not statistically significant ($p > 0.05$). There is a problem regarding `Long-term(Abel)` in the case of $|time - time(t)| > max_{time} - min_{time}$, i.e., $\frac{|time - time(t)|}{max_{time} - min_{time}} > 1$. In this case, we can observe that the weight is increasing with a higher value of *time*, which should be decreased instead since a higher value of *time* denotes t is becoming older than before.

It is worth noting that, although recent interests of users are more important than older ones, we still need to include older history to retrieve the most complete profiles of user interests. As we can see from the figure, `Long-term(Ahmed`α`)` which slows down the decay of user interests, outperforms `Long-term(Ahmed)` consistently in terms of all evaluation methods, which shows the slow decay of user interests on Twitter. Similar conclusions were reached in [20] based on a user study. Orlandi et al. [20] showed in their experiment that, by setting $\beta = 360days$ in Equation 2 leads to better performance compared to setting $\beta = 120days$. Note

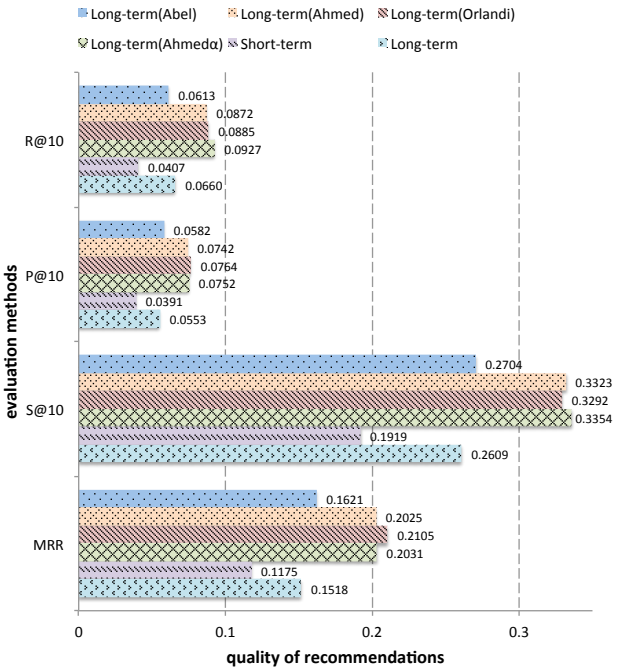


Figure 3: The quality of recommendations with different methods considering dynamics of user interests

that, by setting β to a larger constant in Equation 2 (`Long-term(Orlandi)`) as well as defining a longer period for each abstraction in Equation 3 (μ_{week}, μ_{month}), we are slowing down the decay of older interests of users. The results based on different parameters of Equation 2 and 3 indicate that, the quality of user modeling increases by giving a higher weight to the recent interests of users but decreases when the weight of recent interests is too high. In other words, we still need to include older history for building user interest profiles.

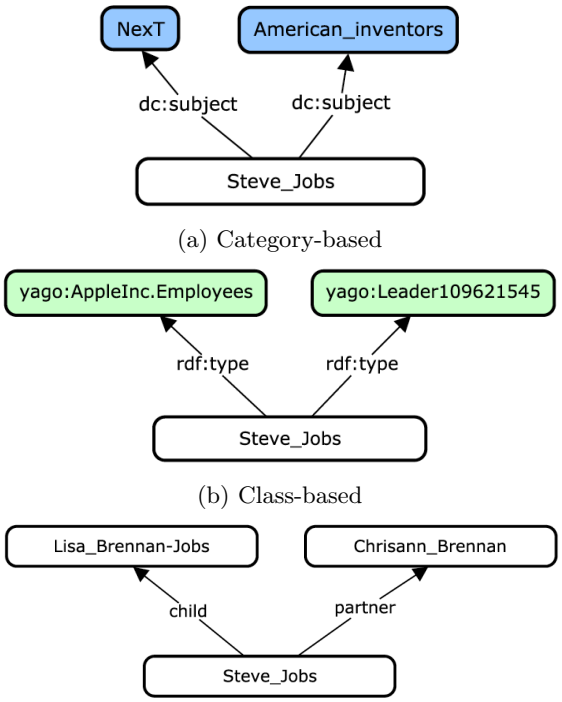
5.3 Leveraging DBpedia for user modeling

Where previous works mainly focused on category-based extension strategy using DBpedia, we investigate three different types of core extension strategies for *primitive interests* as below and the combination of these core strategies.

- **Category-based**: the strategy extends *primitive interests* using their category information (Figure 4(a)), which relies on the category system of Wikipedia[8] to capture the idea of a "theme", i.e., a subject of the entity [16].
- **Class-based**: the strategy extends *primitive interests* using their class information (Figure 4(b)), which is provided via `rdf:type` statements for all DBpedia entities using classification from YAGO [26].
- **Property-based**: the method extends *primitive interests* with connected entities via various properties defined in DBpedia Ontology (Figure 4(c)).

[7] https://yahoo.com/

[8] https://en.wikipedia.org/wiki/Main_Page



(a) Category-based

(b) Class-based

(c) Property-based

Figure 4: Three core strategies using DBpedia for extending user interests

As previous studies [20,22] showed that a discounting strategy is required for the extended concepts based on primitive interests, we adopt the discounting strategy in [22]. The extended categories from DBpedia based on primitive interests are discounted as follows:

$$CategoryDiscount = \frac{1}{\alpha} \times \frac{1}{\log(SP)} \times \frac{1}{\log(SC)} \quad (5)$$

where: *SP = Set of Pages belonging to the Category, SC = Set of Sub-Categories*. We set the parameter $\alpha = 2$ as in [22]. Thus, an extended category is discounted heavily if it is a general one (i.e., the category has a great number of pages or sub-categories). In addition, the parameter α denotes the discount of the extended category-based user profiles for combining the entity-based and category-based user profiles.

In the same way, the extended classes using class-based extension strategy can be discounted as follows:

$$ClassDiscount = \frac{1}{\alpha} \times \frac{1}{\log(SP')} \times \frac{1}{\log(SC')} \quad (6)$$

where: *SP' = Set of Pages belonging to the Class, SC' = Set of Sub-Classes*.

In terms of property-based extension strategy, extended entities via different properties are discounted based on the occurrence frequency of a specific property in DBpedia.

$$PropertyDiscount = \frac{1}{\alpha} \times \frac{1}{\log(P)} \quad (7)$$

where: *P = the number of occurrences of a property in the whole DBpedia graph*. The intuition behind *PropertyDiscount* is that entities extended via a property appearing rarely in the DBpedia graph should be given a higher weight than ones extended via a property appearing frequently [23].

One of the benefits of property-based extension strategy is that this strategy strengthens the IDF value of a concept in the CF-IDF weighting scheme as the indirect mentions of the concept by users could be counted. For example, the concept `dbpedia:Montana` has appeared 36 times (which is the Document Frequency of the concept) before applying the extension strategy. However, we observe that the number has increased to 48 after applying the extension strategy as some users indirectly mentioned the topic (e.g., `dbpedia:Virginia_City,_Montana`→`dbpedia-owl:isPartOf`→`dbpedia:Montana`).

Figure 5 presents the number of distinct concepts in user profiles after applying the three different extension strategies. As we can see from the figure, *category-based* extension strategy reveals more information (i.e., a greater number of concepts) in comparison to *class-* and *property-based* extension strategies. On average, *entity-based* user profiles have 224 concepts before any extension. After applying *category-*, *class-* and *property-based* extension strategies, the numbers of concepts in user interest profiles are increased to 1,865, 1,317 and 1,152, respectively.

5.3.1 Results

The quality of recommendations using different extending strategies is summarized in Table 2. While there is no significant difference between core strategies, *category-based* extension strategy achieves the best performance in terms of MRR while *property-based* extension strategy achieves the best performance in terms of S@10, R@10 and P@10.

The results presented in Table 2 also reveal that the combination of different extending strategies for inferring user interests further enhances the quality of user modeling in the context of link recommendations. The *category & class & property-based* extension strategy provides the best performance in terms of MRR and improves the performance of recommendations significantly compared to the *class-based* extension strategy. Regarding other evaluation metrics, we observe that *category & property-based* extension strategy provides the best performance compared to other core ex-

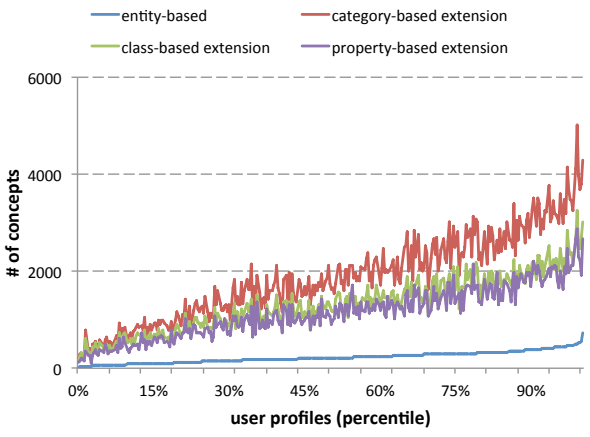


Figure 5: The number of concepts after extending user interest profiles with different core strategies

Table 2: Overview of the different strategies for extending user profiles with background knowledge from DBpedia

extension strategy	MRR	S@10	R@10	P@10
core strategies:				
category-based	**0.2044**	0.3447	0.0928	0.0798
class-based	0.1939	0.3261	0.0861	0.0752
property-based	0.2017	**0.3478**	**0.0956**	**0.0804**
combined strategies:				
category & class-based	0.2065	0.3416	0.0914	0.0780
category & property-based	0.2083	**0.3540**	**0.0993**	**0.0820**
class & property-based	0.2063	0.3478	0.0896	0.0786
category & class & property-based	**0.2103**	0.3478	0.0947	0.0811

tension strategies as well as other combined strategies. The results imply that extension strategies based on different types of information from DBpedia complement each other and the combination of these types of information can improve the quality of user modeling further.

5.4 Comparison with baseline methods

To evaluate our user modeling strategies, we compare our strategies with two baseline methods. In the following, let $um(param1, param2, param3)$ denote a user modeling strategy where *param1*, *param2* and *param3* represent *weighting scheme*, *interest decay function* and *extension strategy*, respectively. The first baseline method from [22] then can be described as $um(CF, none, category)$, which uses CF as the weighting scheme and enriches user profiles using related categories. The second baseline $um(CF, Orlandi, none)$ is entity-based user profiles with an interest decay function from [20]. Our user modeling strategies can be described in the same way. For instance, $um(CF\text{-}IDF, Ahmed\alpha, none)$ denotes the user modeling strategy using CF-IDF as the weighting scheme and using $Ahmed\alpha$ as the interest decay function. For brevity we only present the results for our user modeling strategies $um(CF\text{-}IDF, Ahmed\alpha, none)$, $um(CF\text{-}IDF, none, category \ \& \ property)$ and $um(CF\text{-}IDF, Ahmed\alpha, category \ \& \ property)$ compared to the baseline methods (see Figure 6).

As we can see from the figure, our user modeling strategies clearly outperform baseline methods significantly. Overall, user profiles with the user modeling strategy $um(CF\text{-}IDF, none, category \ \& \ property)$ which exploits semantic information from DBpedia provide the best performance in terms of all evaluation metrics.

Figure 6 further illustrates the results of our user modeling strategies combining the two dimensions (i.e., dynamics and semantics of user interests). Although we expect combing two dimensions would further improve the quality of user modeling, as we can see from the figure, $um(CF\text{-}IDF, Ahmed\alpha, category \ \& \ property)$ does not further improve the quality of user modeling compared to using a single dimension, i.e., $um(CF\text{-}IDF, Ahmed\alpha, none)$ and $um(CF\text{-}IDF, none, category \ \& \ property)$. Further research is required to understand to what extent can different dimensions be exploited together for improving user modeling. Similar results can be observed in combinations of other methods (e.g., $um(CF\text{-}IDF, Orlandi, none)$ and $um(CF\text{-}IDF, none, category \ \& \ property)$).

6. CONCLUSIONS

The paper explored two dimensions of user modeling: (1) dynamics of user interests, and (2) extending strategy leveraging different types of information from DBpedia for user interests. First, we evaluated our choice for using CF-IDF as our weighting scheme for user modeling. On top of that, we conducted a comparative study on different user modeling approaches incorporating the dynamics of user interests in the literature. Furthermore, we looked at different core strategies for extending user interest profiles using different types of information from DBpedia (i.e., *categories, classes and connected entities via different properties*) as well as mixed approaches using these core strategies. The results suggested that mixed approaches can provide the best performance compared to using a single type of information for extending user profiles. Finally, we evaluated our user

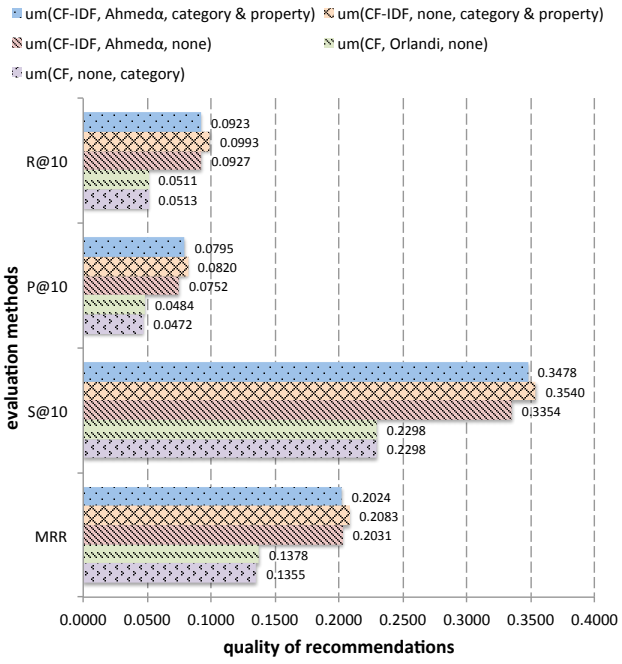


Figure 6: Comparison between our user modeling strategies and baseline methods

modeling strategies and showed that our strategies outperform two baseline approaches significantly in the context of link recommendations. In our forthcoming research [21], we would deepen the investigation of how to combine different dimensions of user modeling to improve the quality of user interest profiles.

7. 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).

8. REFERENCES

[1] A. Abdel-Hafez and Y. Xu. A survey of user modelling in social media websites. *Computer and Information Science*, 6(4):p59, 2013.

[2] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web. In *Proceedings of the 3rd International Web Science Conference*, page 2. ACM, 2011.

[3] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing user modeling on twitter for personalized news recommendations. In *User Modeling, Adaption and Personalization*, pages 1–12. Springer, 2011.

[4] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In *The Semanic Web: Research and Applications*, pages 375–389. Springer, 2011.

[5] F. Abel, C. Hauff, G.-J. Houben, and K. Tao. Leveraging User Modeling on the Social Web with Linked Data. In *Web Engineering SE - 31*, pages 378–385. Springer, 2012.

[6] A. Ahmed, Y. Low, M. Aly, V. Josifovski, and A. J. Smola. Scalable distributed inference of dynamic user interests for behavioral targeting. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 114–122. ACM, 2011.

[7] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[8] C. Budak, A. Kannan, R. Agrawal, and J. Pedersen. Inferring user interests from microblogs. Technical report, 2014.

[9] D. Godoy and A. Amandi. Modeling user interests by conceptual clustering. *Information Systems*, 31(4-5):247–265, jun 2006.

[10] F. Goossen, W. IJntema, F. Frasincar, F. Hogenboom, and U. Kaymak. News personalization using the CF-IDF semantic recommender. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, page 10. ACM, 2011.

[11] M. Harvey, F. Crestani, and M. J. Carman. Building User Profiles from Topic Models for Personalised Search. *Cikm*, pages 2309–2314, 2013.

[12] T. Heath and C. Bizer. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136, 2011.

[13] P. Jain, P. Kumaraguru, and A. Joshi. @i seek 'fb.me': identifying users across multiple online social networks. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1259–1268. ACM, 2013.

[14] P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth. User Interests Identification on Twitter Using a Hierarchical Knowledge Base. In *The Semantic Web: Trends and Challenges*, pages 99–113. Springer, 2014.

[15] S. Kinsella, M. Wang, J. G. Breslin, and C. Hayes. Improving categorisation in social media using hyperlinks to structured data sources. In *The Semanic Web: Research and Applications*, pages 390–404. Springer, 2011.

[16] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, and S. Auer. Dbpedia-a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2013.

[17] C. Lu, W. Lam, and Y. Zhang. Twitter user modeling and tweets recommendation based on wikipedia concept graph. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[18] M. Michelson and S. A. Macskassy. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, pages 73–80. ACM, 2010.

[19] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. 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*, pages 251–260. ACM, 2010.

[20] F. Orlandi, J. Breslin, and A. Passant. Aggregated, interoperable and multi-domain user profiles for the social web. In *Proceedings of the 8th International Conference on Semantic Systems*, pages 41–48. ACM, 2012.

[21] G. Piao. Towards Comprehensive User Modeling on the Social Web for Personalized Recommendations. In *User Modeling, Adaptation, and Personalization*. ACM, 2016.

[22] G. Piao and J. G. Breslin. Analyzing Aggregated Semantics-enabled User Modeling on Google+ and Twitter for Personalized Link Recommendations. In *User Modeling, Adaptation, and Personalization*. ACM, 2016.

[23] G. Piao and J. G. Breslin. Measuring Semantic Distance for Linked Open Data-enabled Recommender Systems. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 315–320. ACM, 2016.

[24] K. Ramanathan and K. Kapoor. Creating User Profiles Using Wikipedia. In *Conceptual Modeling - ER 2009*, volume 5829, chapter 31, pages 415–427. Springer Berlin Heidelberg, 2009.

[25] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. 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*, pages 841–842. ACM, 2010.

[26] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM, 2007.