Interest Representation, Enrichment, Dynamics, and Propagation: A Study of the Synergetic Effect of Different User Modeling Dimensions for Personalized Recommendations on Twitter

Guangyuan Piao and John G. Breslin

Insight Centre for Data Analytics
National University of Ireland Galway
IDA Business Park, Lower Dangan, Galway, Ireland
{guangyuan.piao}@insight-centre.org
{john.breslin}@nuigalway.ie

Abstract. Microblogging services such as Twitter have been widely adopted due to the highly social nature of interactions they have facilitated. With the rich information generated by users on these services, user modeling aims to acquire knowledge about a user's interests, which is a fundamental step towards personalization as well as recommendations. To this end, researchers have explored different dimensions such as (1) *Interest Representation*, (2) *Content Enrichment*, (3) *Temporal Dynamics* of user interests, and (4) *Interest Propagation* using semantic information from a knowledge base such as DBpedia. However, those dimensions of user modeling have largely been studied separately, and there is a lack of research on the synergetic effect of those dimensions for user modeling. In this paper, we address this research gap by investigating 16 different user modeling strategies produced by various combinations of those dimensions. Different user modeling strategies are evaluated in the context of a personalized link recommender system on Twitter. Results show that *Interest Representation* and *Content Enrichment* play crucial roles in user modeling, followed by *Temporal Dynamics*. The user modeling strategy considering *Interest Representation*, *Content Enrichment* and *Temporal Dynamics* provides the best performance among the 16 strategies. On the other hand, *Interest Propagation* has little effect on user modeling, even when leveraging a rich *Interest Representation* or considering *Content Enrichment*.

Keywords: User modeling, Personalization, Twitter, DBpedia

1 Introduction

With the popularity of microblogging services such as Twitter[1], the amount of information available on the Social Web is increasing exponentially. While this

[1] https://www.twitter.com

Table 1: A sample tweet posted by Bob

My Top 3 #lastfm Artists: Eagles of Death Metal(14), The Black Keys(6) & The Wombats(6). http://www.last.fm/user/bob

information is a valuable resource, the sheer volume limits its value [9]. On the Social Web, as the amount of information available causes information overload for users, the demand for personalized approaches towards information consumption increases. User (interest) modeling aims to analyze user activities on the Social Web in order to provide personalized services for users. To create qualitative and quantitative user models for microblogging services such as Twitter, several design dimensions have been investigated in previous studies.

Interest Representation. Defining the representation of user interests is a fundamental step for user modeling. Several approaches such as *bag-of-words*, *topic models* and *bag-of-concepts* have been used for representing user interests. For example, the bag-of-concepts approach uses concepts for representing user interests. Given a tweet posted by a user named Bob (see Table 1), we can assume that the user is interested in DBpedia[2] entities such as `dbpedia`[3]`:The_Black_Keys` and `dbpedia:The_Wombats`. This approach has been adopted in recent studies, where background knowledge on concepts from a knowledge base (KB) (defined as the combination of an ontology and instances of the classes in the ontology [22]) can be leveraged for extending user interests. Throughout the paper, by a *concept* we mean an *entity*, *category* or *class* from a KB (e.g., DBpedia) for representing user interests.

Content Enrichment. As the user-generated content (UGC) on microblogging services is short in nature (e.g., 140 characters for tweets), there is a need to enrich this short content to better understand the context of it. Embedded links (URLs) in a tweet can be used to enrich the short content and provide additional information about the tweet. For example, we can follow the link in the sample tweet to retrieve more information about Bob's musical interests. Many sources have shown that a large portion of tweets and retweets contain links[4][5].

Temporal Dynamics. The interests of users can change over time. To capture the dynamics of user interests, some previous studies have used short-term profiles (e.g., considering a user's activities during the last two weeks alone), while others have proposed interest decay functions to discount older interests.

Interest Propagation. On top of the concept-based representation of user interests, researchers have further exploited semantic information from a knowledge base, which provides cross-domain background knowledge about concepts. For instance, DBpedia is the semantic representation of Wikipedia[6], and is a

[2] `http://wiki.dbpedia.org`
[3] The prefix `dbpedia` denotes `http://dbpedia.org/resource/`
[4] `http://marketingrelevance.com/news/04/tweet-interesting-information/`
[5] `http://goo.gl/RGC16n`
[6] `https://en.wikipedia.org/wiki/Main_Page`

KB consisting of a large amount of facts about concepts. This can be helpful for extending user interest profiles in order to provide more semantic information about users. For instance, from the tweet in Table 1, we can further infer that the user is interested in `dbpedia:Indie_rock` as both `dbpedia:The_Black_Keys` and `dbpedia:The_Wombats` are pointing to `dbpedia:Indie_rock` via the property `dbpedia-owl`[7]`:genre`. Throughout the paper, we denote the concepts that can be directly extracted from a user's tweets as ***primitive interests*** (e.g., `dbpedia:The_Wombats`), and the concepts that can be propagated from those primitive interests as ***propagated interests*** (e.g., `dbpedia:Indie_rock`).

Although related work reveals many promising insights with respect to those user modeling dimensions, there exists little research on studying the synergetic effect achieved by considering those dimensions together. As those dimensions are not necessarily exclusive of each other, this has in turn motivated us to implement a user modeling framework which can exploit different dimensions at the same time for generating user interest profiles. We then evaluate different user interest profiles generated by different user modeling strategies in the context of a personalized link (URL) recommender system on Twitter.

The contributions of this work are summarized as follows.

- We implemented a user modeling framework, which can incorporate different combinations of four dimensions: (1) *Interest Representation*, (2) *Content Enrichment*, (3) *Temporal Dynamics*, and (4) *Interest Propagation*, to investigate (how) can we combine these different dimensions to retrieve better user interest profiles. To our knowledge, this is the first comprehensive study on these four dimensions.

- We evaluate 16 user modeling strategies generated by different combinations of methods for those four dimensions in the context of link recommendations on Twitter using four different evaluation metrics.

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. Experiment results are presented in Section 5. Finally, Section 6 concludes the paper with some future work.

2 Related Work

In this section, we provide an overview of some related work from the literature for each of the aforementioned dimensions in user modeling.

Representation of user interests. To represent user interest profiles, researchers began by using *word-based* approaches such as *bag-of-words* [8,17] and *topic modeling* [10]. Degemmis et al. [8] proposed a specific *word-based* approach - using WordNet[8] synsets (which are unordered sets of synonyms) to represent

[7] The prefix `dbpedia-owl` denotes `http://dbpedia.org/ontology/`
[8] `https://wordnet.princeton.edu/`

user interests. They showed that their *bag-of-synsets* approach outperformed a *bag-of-words* approach. As *word-based* approaches focus on the words themselves and do not provide semantic information about the words or the relationships among them, a research direction has been proposed over the past few years that uses *concept-based* representations of user interests using a KB in Linked Data form (e.g., Freebase, DBpedia) [4, 5, 19, 21] or using an encyclopedia such as Wikipedia [12, 15, 16, 18].

Enrichment for short messages. The length of posts on microblogging services such as Twitter is usually short, which makes it difficult to detect the semantics of these messages. Researchers [4,13] have used the content of embedded links (URLs) in short messages to enrich the content. For example, Abel et al. [4] exploited URLs shared via tweets, and devised a methodology to link tweets to news articles in their monitored news pool so as to use the content of news articles to enrich the user interest profiles for their news recommender system. They showed that enriching short content for retrieving user interests enhances the variety and quality of the generated user profiles.

Dynamics of user interests. Many methods have been proposed to incorporate the temporal dynamics of user interests based on the hypothesis that the interests of users change over time [2,3,7,19]. For example, Abel et al. [3] studied 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 their entire historical UGC. Another line of work [2,7,19] that incorporates temporal dynamics applies a decay function to the interests of users. The rationale behind the decay function is that higher weights should be given to interests that have occurred recently and lower weights given to older interests.

Interest propagation using background knowledge. Based on *concept-based* user profiles, researchers have also proposed using the rich semantic information from a KB to extend the interests of users. Orlandi et al. [19] proposed *category-based* user profiles based on the category information of entities from DBpedia. As well as proposing a straightforward extension that gives equal weight to each extended category with respect to an entity, they also proposed a discounting strategy for those extended categories. Piao et al. [20] proposed a mixed approach that combines the entity- and category-based profiles with the discounting strategy from [19], and proved that the mixed approach performs better than either the entity- or category-based approach. Building on this in a later work [21], the authors showed that by using Concept Frequency-Inverse Document Frequency (CF-IDF) as the weighting scheme and by leveraging different types of information from DBpedia to extend user profiles (i.e., *categories, and connected entities via different properties*), the quality of user modeling can be improved.

There are also some studies looking at user modeling for a specific domain of user interests. For example, Abel et al. [5] proposed using DBpedia to extend user profiles with respect to point of interests (POI), and Nishioka et al. [18] explored different factors for modeling user interests with respect to scientific

publications in the economic domain. Differing 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.

While related work reveals several insights regarding each dimension of user modeling, hybrid approaches combining those different dimensions are considered only to a limited degree. For example, after enriching tweets with the content of embedded links, it would be interesting to explore if interest propagation using background knowledge further improves the quality of user modeling, or if it has little or no effect since enough information may already be available from a user's primitive interests.

3 Content-based User Modeling

In this section, we first introduce user interest profiles as defined in our work, and then present a general process for generating user interest profiles (Section 3.1). Subsequently, we provide details of the methods for each of the user modeling dimensions used in the process (Section 3.2).

In this work, we use DBpedia concepts or WordNet synsets to represent the interests of users. The generic model for representing the interest profiles of users is specified as follows.

Definition 1. *The interest profile of a user $u \in U$ is a set of weighted DBpedia concepts or WordNet synsets, where, with respect to a given user u who has an interest $i \in I$, its weight $w(u, i)$ is computed by a certain function w.*

Here, U denotes the set of users, and I denotes the set of concepts in DBpedia and synsets in WordNet. The weighting scheme $w(u, i)$ measures the importance of a concept with respect to a user. Previous studies showed that using CF-IDF as the weighting scheme provides better performance than using a Concept Frequency (CF) weighting scheme for user modeling in the context of recommender systems [18,21]. 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 have a higher weight. In the same way, we use Interest Frequency-Inverse Document Frequency (IF-IDF) as the weighting scheme for our experiments. More formally, it is defined as follows.

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

where M is the total number of users, and m_i is the number of users interested in a concept/synset i.

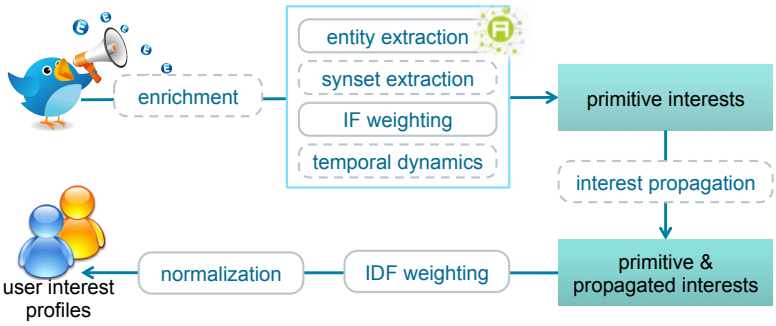


Fig. 1: The process of generating user interest profiles on Twitter

3.1 The process of generating user interest profiles

Figure 1 presents the process of generating user interest profiles for Twitter considering the aforementioned four different user modeling dimensions. The components with dotted lines are options that can be either "enabled" or "disabled" for this user modeling. The process has three major steps:

(1) Primitive interests extraction. For a given user, we extract all *primitive interests* (DBpedia entities or WordNet synsets) within the UGC of a user. If the *enrichment* component is enabled, the content of links embedded in the UGC will also be used for extracting primitive interests.

- Entities are extracted using the `Aylien API`[9]. 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". Interest Frequency (IF) is applied to denote the importance of a concept with respect to a user. In addition, it might adhere to strategies for incorporating the *temporal dynamics* of user interests.
- WordNet synsets can be extracted at the same time as extracting entities. The rationale behinds this is that syntactic information can complement semantic information for generating user interest profiles. For example, given a tweet: "*Just completed a 3.89 km ride. We're gonna need more...*", we can extract synsets such as : s1=[kilometer, kilometre, km, klick (a metric unit of length equal to 1000 meters (or 0.621371 miles))] and s2 = [drive, ride (a journey in a vehicle (usually an automobile))], which denote the user interests that would be missed if we used a concepts-only approach.

(2) Interest propagation. This component can apply propagation strategies to primitive interests based on background knowledge from DBpedia. The

[9] `http://aylien.com`

Table 2: The design space of user modeling, spanning 2x2x2x2=16 possible user modeling strategies.

User Modeling Dimensions	Interest Representation	Content Enrichment	Temporal Dynamics	Interest Propagation
Options	*concept*	*enabled*	*enabled*	*enabled*
	synset & concept	*disabled*	*disabled*	*disabled*

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) to the user interest profile, and further normalizes the profile so that the sum of all weights in the profile is equal to 1: $\sum_{i \in I} w(u,i) = 1$.

Based on the optional components for user modeling (shown with dotted lines in Figure 1), there are 16 possible strategies which are displayed in Table 2. In the following subsection, we provide details of the methods for each dimension.

3.2 Methods for each dimension

Interest Representation: (1) concept, or (2) synset & concept. *Entity recognition* and *synsets extraction* are performed in the first step to extract *primitive interests* from a user's tweets.

Entity recognition in tweets is a challenging task due to the informal nature of and ungrammatical language in tweets. Since our focus in this work is on user modeling and not on entity recognition, we have used an existing solution for entity recognition (as does related literature on user modeling). Different NLP APIs have been used for DBpedia/Wikipedia entity recognition in the literature. For example, Kapanipathi et al. [12] used the Zemanta API (which is no longer available) after comparing it to other APIs such as DBpedia Spotlight[10], Fattane et al. [23] used tag.me[11], and Piao et al. [21] used the Aylien API, respectively. To better investigate the performance of different APIs, we used the Twitter dataset from [14] which contains annotated 1,603 tweets in total where 1,233 of them contain Wikipedia entities. We tested three different NLP APIs: the Aylien API, tag.me, and the Alchemy

Table 3: Evaluation of NLP APIs for DBpedia/Wikipedia entity recognition

API	**Precision**	**Recall**	**F-measure**
Aylien	0.27	0.26	0.26
Alchemy	0.21	0.17	0.19
tag.me	0.12	0.15	0.14

[10] `http://spotlight.dbpedia.org/rest/annotate`, the web service was not accessible at the time of writing this paper

[11] `https://tagme.d4science.org/tagme/`

API[12], which all provide functionality for extracting entities from a given text and representing these with their corresponding DBpedia/Wikipedia URIs. A comparative performance is displayed in Table 3. We opted to use the Aylien API for our experiment since (1) it extracts DBpedia entities (*primitive interests*) identified in tweets, and gives their corresponding URIs, (2) it has relatively superior performance to the other APIs as shown in Table 3, and (3) it provides 6,900 calls per day, provided on request for research purposes.

Synset extraction is included in the investigation since not everything will have a corresponding concept that is covered by a KB, especially in the case of Twitter where new concepts/topics emerge everyday. Also knowledge bases such as DBpedia can lack full coverage for the lexicographic senses of lemmas, which can be provided by a lexical database such as WordNet. To this end, we adopt a method from [8] which extracts WordNet synsets to build *synset-based* user interest profiles.

Content Enrichment: (1) enabled, or (2) disabled. We leverage the content of links embedded in a tweet to enrich the original post content. Based on the selected option for the dimension *Interest Representation*, we apply the same extraction method for the content of embedded links. Therefore, in the case of *concepts* being used for *Interest Representation*, the *concepts* extracted from the links embedded in tweets will also be considered as user interests if the *Content Enrichment* dimension option is enabled.

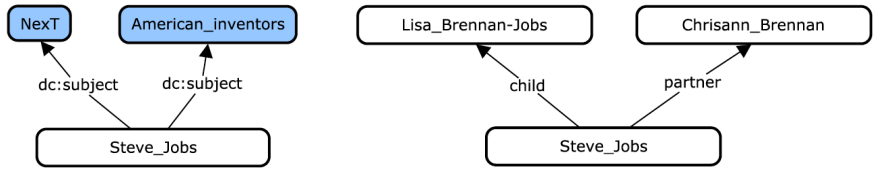
Temporal Dynamics: (1) enabled, or (2) disabled. In [21], the authors conducted a comparative study on different interest decay functions [2, 6, 19] for incorporating the temporal dynamics of user interests in the context of recommender systems on Twitter. Results showed that those functions have similar performance. We choose a variant of the interest decay function from [6], which performed best overall in the comparative study [21]. This decay function [21] measures the expected weight in terms of an interest i for a user k at time t by combining three levels of abstractions, using a weighted sum as below:

$$w_{ki}^{t} = \mu_{2week} w_{ki}^{t,2week} + \mu_{2month} w_{ki}^{t,2month} + \mu_{all} w_{ki}^{t,all} \quad (1)$$

where $\mu_{2week} = \mu$, $\mu_{2month} = \mu^2$ and $\mu_{all} = \mu^3$ and $\mu \in [0, 1]$. We set μ as e^{-1}, in the same manner as [6, 21], for our experiment.

Interest Propagation: (1) enabled, or (2) disabled. In [21], the authors also proposed different propagation strategies exploiting different types of background knowledge from DBpedia. Overall, the propagation strategy extending *primitive interests* with categories (Figure 2(a)) and entities connected via different properties (Figure 2(b)) in DBpedia provided the best performance compared to other state-of-art propagation strategies.

[12] `http://www.alchemyapi.com/`



(a) Category-based (b) Property-based

Fig. 2: Three core strategies using DBpedia for extending user interests

As previous studies [19, 20] showed that a discounting strategy is required for the extended concepts based on primitive interests, the authors [21] applied a discounting strategy from [20] for the extended categories as follows:

$$CategoryDiscount = \frac{1}{\alpha} \times \frac{1}{\log(SP)} \times \frac{1}{\log(SC)} \tag{2}$$

where: *SP = Set of Pages belonging to the Category, SC = Set of Sub-Categories.* We set the parameter $\alpha = 2$ as in [21]. 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 propagated interests from primitive interests. Regarding the property-based extension strategy (Figure 2(b)), extended entities via different properties are discounted based on the occurrence frequency of a specific property in DBpedia [21]:

$$PropertyDiscount = \frac{1}{\alpha} \times \frac{1}{\log(P)} \tag{3}$$

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.

4 Experiment Setup

In the following section, we describe the Twitter dataset used in our experiment (Section 4.1), and the evaluation methodology (Section 4.2). Subsequently, we present the results using 16 different user modeling strategies in the context of link recommendations on Twitter (Section 4.3).

4.1 Twitter dataset

We used the Twitter dataset from [20]. The dataset contains all tweets published by 480 *active* users on Twitter (a user is *active* if the user published at least 100 posts [11, 15, 20]). The main details of the dataset are presented in Table 4.

Table 4: Twitter 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

Dataset for link recommendations. In the same way as [21], we further selected users who shared at least one link (URL) in their tweets during the previous two weeks. We only consider links that have at least four topics (concepts) to filter out non-topical links (e.g., links sharing a current location via Swarm[13]). 322 out of 480 users met this criteria, publishing 247,676 tweets in total.

4.2 Evaluation methodology

One thing to note is that our main goal is to analyze and compare the applicability of 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. Therefore, we apply a lightweight content-based algorithm, similar to the one used in [5], that recommends items according to their *cosine* similarity with a given user profile.

Definition 2. *Recommendation Algorithm: given a user profile P_u and a set of candidate links $N = \{P_{i1}, ..., P_{in}\}$, 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.*

We assumed a user was interested in the content of a link (URL) if the link was shared via the user's tweets. The ground truth of links was a set of links shared via the user's tweets within the last two weeks. To construct candidate links for recommendations, we used the ground truth links from 322 users, as well as the links shared by other users but not shared by the 322 users in the dataset. 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 a user interest profile, the recommendation system then recommends the top-N links from the candidate links that the user might be interested in, using the recommendation algorithm (see Definition 2). The quality of the top-N recommendations was measured via the following metrics, which had been used in previous studies [3, 5, 19, 21].

[13] `https://www.swarmapp.com`

- **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 ranked.
- **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.

We focused on $N = 10$ as our recommendation system lists 10 link recommendations to a user. We used the *bootstrapped paired t-test*[14] (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 Results

In this section, we present the results of experiments using different user modeling strategies in the context of link recommendations. In the following, let *um*(*representation*, *enrichment*, *dynamics*, *semantics*) denote a user modeling strategy where four parameters: *representation*, *enrichment*, *dynamics* and *semantics* represent the four dimensions *Interest Representation*, *Content Enrichment*, *Temporal Dynamics* and *Interest Propagation*, respectively. We use "none" to denote when a certain dimension is disabled. For instance, *um*(*concept*, *none*, *none*, *none*) denotes a user modeling strategy using concepts for *Interest Representation* without considering any other dimensions. *um*(*synset* & *concept*, *enrichment*, *none*, *none*) denotes a user modeling strategy using synsets and concepts for *Interest Representation*, and tweets are enriched with the content of embedded links when extracting user interests (i.e., the dimension *Content Enrichment* is enabled).

Table 5 summarizes the recommendation performance using the 16 user modeling strategies in terms of different evaluation metrics. The results are sorted in descending order in terms of MRR. Overall, the best performing strategy is *um*(*synset* & *concept*, *enrichment*, *dynamics*, *none*), which uses DBpedia concepts and WordNet synsets for *Interest Representation*, and considers all dimensions except *Interest Propagation*. Table 5 shows the importance of (1) *Content Enrichment*, and (2) *Interest Representation* in user modeling. For instance, the strategies enriching tweets with embedded links (1-8 in Table 5) clearly have better performance than the ones without any enrichment (9-16), using the same option for *Interest Representation*. In terms of *Interest Representation* with or without *Content Enrichment*, we observe that using DBpedia concepts with WordNet synsets (1-4 and 9-12) always provides better performance than using concepts alone (5-8 and 13-16). This indicates that exploiting

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

Table 5: Performance of link recommendations using 16 user modeling strategies four different evaluation metrics. The results are sorted in descending order in terms of MRR.

	User Modeling Strategies	MRR	S@10	R@10	P@10
1.	um(synset & concept, enrichment, dynamics, none)	0.3251	0.5062	0.1700	0.1304
2.	um(synset & concept, enrichment, dynamics, propagation)	0.3198	0.4938	0.1654	0.1298
3.	um(synset & concept, enrichment, none, none)	0.3146	0.4876	0.1595	0.1286
4.	um(synset & concept, enrichment, none, propagation)	0.3107	0.4752	0.1534	0.1267
5.	um(concept, enrichment, dynamics, none)	0.2942	0.4193	0.1405	0.1047
6.	um(concept, enrichment, none, none)	0.2886	0.4379	0.1392	0.1062
7.	um(concept, enrichment, dynamics, propagation)	0.2802	0.3975	0.1287	0.0988
8.	um(concept, enrichment, none, propagation)	0.2736	0.4130	0.1332	0.1006
9.	um(synset & concept, none, dynamics, none)	0.2511	0.4255	0.1257	0.0988
10.	um(synset & concept, none, dynamics, propagation)	0.2502	0.4193	0.1259	0.0997
11.	um(synset & concept, none, none, none)	0.2436	0.4068	0.1231	0.0978
12.	um(synset & concept, none, none, propagation)	0.2386	0.3913	0.1179	0.0984
13.	um(concept, none, none, propagation)	0.2083	0.3540	0.0993	0.0820
14.	um(concept, none, dynamics, none)	0.2031	0.3354	0.0927	0.0752
15.	um(concept, none, dynamics, propagation)	0.2024	0.3478	0.0923	0.0795
16.	um(concept, none none, none)	0.1518	0.2609	0.0660	0.0553

semantic and lexical knowledge from DBpedia as well as WordNet for *Interest Representation* improves the quality of user modeling.

Table 6 further illustrates statistical differences between the 16 user modeling strategies in terms of MRR. Overall, the results of other evaluation metrics are similar to the MRR and thus omitted for reasons of brevity. The vertical and horizontal dimensions of the table show a comparison between the 16 strategies. As we can see from the table, there are various significant differences between the strategies ($p < .05$, marked in bold font). For example, strategies using concepts and synsets for the dimension *Interest Representation* always significantly outperform strategies using concepts, when other dimensions are kept the same (e.g., 1 and 5). The dimension *Interest Propagation* plays an important role when we use concepts for *Interest Representation* without *Content Enrichment* (13-16). However, when we have a rich interest representation (i.e., using concepts and synsets together) or rich content by enrichment, *Interest Propagation* has little effect on the quality of user modeling, i.e., there is no statistical difference between a user modeling strategy with *Interest Propagation* and one without any propagation (1-12). One of the possible reasons might be the rich interest representation, and content is giving sufficient knowledge about user interests. Additionally, the "insufficient quality" of extracted DBpedia entities from tweets using APIs (see the precision in Table 3 in Section 3.2), could result

Table 6: Results of p-values over the 16 user modeling strategies in terms of link recommendations on Twitter (marked in bold font if $p < .05$). Strategies are sorted by MRR results as shown in Table 5.

		2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.
1.	um(synset & concept, enrichment, dynamics, none)	.14	.17	.11	**.01**	**.02**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**
2.	um(synset & concept, enrichment, dynamics, propagation)		.35	.21	**.04**	**.04**	**.01**	**.01**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**
3.	um(synset & concept, enrichment, none, none)			.24	.10	.05	**.03**	**.01**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**
4.	um(synset & concept, enrichment, none, propagation)				.18	.10	**.03**	**.02**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**
5.	um(concept, enrichment, dynamics, none)					.31	.05	**.03**	**.02**	**.02**	**.01**	**.01**	**.00**	**.00**	**.00**	**.00**
6.	um(concept, enrichment, none, none)						.26	.05	**.03**	**.02**	**.01**	**.01**	**.00**	**.00**	**.00**	**.00**
7.	um(concept, enrichment, dynamics, propagation)							.26	.10	.08	.05	**.03**	**.00**	**.00**	**.00**	**.00**
8.	um(concept, enrichment, none, propagation)								.13	.13	.07	**.04**	**.00**	**.00**	**.00**	**.00**
9.	um(synset & concept, none, dynamics, none)									.42	.20	.08	**.01**	**.00**	**.00**	**.00**
10.	um(synset & concept, none, dynamics, propagation)										.22	.08	**.01**	**.01**	**.00**	**.00**
11.	um(synset & concept, none, none, none)											.15	**.02**	**.01**	**.01**	**.00**
12.	um(synset & concept, none, none, propagation)												**.04**	**.03**	**.02**	**.00**
13.	um(concept, none, none, propagation)													.32	.27	**.00**
14.	um(concept, none, dynamics, none)														.46	**.00**
15.	um(concept, none, dynamics, propagation)															**.00**
16.	um(concept, none, none, none)															

in inaccurate interest propagation based on incorrect entities. This might limit the contribution of propagated interests towards user modeling.

Similar results can be found for temporal dynamics. Although considering *Temporal Dynamics* increases the performance significantly when we use concepts for *Interest Representation* without *Content Enrichment* (13-16), there is no significant difference between strategies with a rich interest representation

and rich content (1-12). Nevertheless, we observe that in all of the cases using concepts and synsets for *Interest Representation*, considering the *Temporal Dynamics* dimension provides the best performance (see 1, 9 in Table 5).

To sum up, the two dimensions *Interest Representation* and *Content Enrichment* play significant roles in user modeling, followed by *Temporal Dynamics*. Although the contribution of content enrichment via embedded links might depend on the percentage of embedded links, it is an important and valuable source for enrichment as a large number of tweets are posted with links[15]. The results also show that the *Interest Propagation* dimension had little effect on user modeling when considering different dimensions together, which is different from previous studies considering one or two dimensions [2, 19–21].

6 Conclusions

In this paper, we investigated different combinations of four dimensions for user modeling on Twitter: (1) *Interest Representation*, (2) *Content Enrichment*, (3) *Temporal Dynamics of user interests*, and (4) *Interest Propagation*, which have not been studied together. As a result, we end up with 16 different user modeling strategies for all possible combinations (see Table 2). These strategies were evaluated in the context of link recommendations on Twitter. The best-performing strategy is *um*(*synset & concept*, *enrichment*, *dynamics*, *none*), which uses DBpedia concepts and WordNet synsets for *Interest Representation* considering *Temporal Dynamics*, with *Content Enrichment*. The results also indicate that *Interest Representation* and *Content Enrichment* are the most important dimensions compared to other dimensions. In future research, we would like to further investigate how different percentages of links in tweets affect the quality of user modeling.

References

1. Abdel-Hafez, A., Xu, Y.: A survey of user modelling in social media websites. Computer and Information Science 6(4), p59 (2013)
2. Abel, F., Gao, Q., Houben, G.J., Tao, K.: Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web. In: Proceedings of the 3rd International Web Science Conference. p. 2. ACM (2011)
3. Abel, F., Gao, Q., Houben, G.J., Tao, K.: Analyzing user modeling on twitter for personalized news recommendations. In: User Modeling, Adaption and Personalization, pp. 1–12. Springer (2011)
4. Abel, F., Gao, Q., Houben, G.J., Tao, K.: Semantic enrichment of twitter posts for user profile construction on the social web. In: The Semanic Web: Research and Applications. pp. 375–389. Springer (2011)
5. Abel, F., Hauff, C., Houben, G.J., Tao, K.: Leveraging User Modeling on the Social Web with Linked Data. In: Web Engineering SE - 31, pp. 378–385. Springer (2012)

[15] 70% of 1 million tweets shared from the US West Coast included links. `http://tnw.to/s3R2i`

6. Ahmed, A., Low, Y., Aly, M., Josifovski, V., Smola, A.J.: Scalable distributed inference of dynamic user interests for behavioral targeting. In: Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining. pp. 114–122. ACM (2011)
7. Budak, C., Kannan, A., Agrawal, R., Pedersen, J.: Inferring user interests from microblogs. Tech. rep. (2014)
8. Degemmis, M., Lops, P., Semeraro, G.: A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation. User Modeling and User-Adapted Interaction 17(3), 217–255 (2007)
9. Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A.: User profiles for personalized information access. In: The adaptive web, pp. 54–89. Springer (2007)
10. Harvey, M., Crestani, F., Carman, M.J.: Building User Profiles from Topic Models for Personalised Search. Cikm pp. 2309–2314 (2013)
11. Jain, P., Kumaraguru, P., Joshi, A.: @i seek 'fb.me': identifying users across multiple online social networks. In: Proceedings of the 22nd international conference on World Wide Web companion. pp. 1259–1268. ACM (2013)
12. Kapanipathi, P., Jain, P., Venkataramani, C., Sheth, A.: User Interests Identification on Twitter Using a Hierarchical Knowledge Base. In: The Semantic Web: Trends and Challenges. pp. 99–113. Springer (2014)
13. Kinsella, S., Wang, M., Breslin, J.G., Hayes, C.: Improving categorisation in social media using hyperlinks to structured data sources. In: The Semanic Web: Research and Applications, pp. 390–404. Springer (2011)
14. Locke, B.W.: Named entity recognition: Adapting to microblogging. Ph.D. thesis (2009)
15. Lu, C., Lam, W., Zhang, Y.: Twitter user modeling and tweets recommendation based on wikipedia concept graph. In: Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence (2012)
16. Michelson, M., Macskassy, S.A.: Discovering users' topics of interest on twitter: a first look. In: Proceedings of the fourth workshop on Analytics for noisy unstructured text data. pp. 73–80. ACM (2010)
17. Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: 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. pp. 251–260. ACM (2010)
18. Nishioka, C., Scherp, A.: Profiling vs. Time vs. Content: What Does Matter for Top-k Publication Recommendation Based on Twitter Profiles? In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries. pp. 171–180. JCDL '16, ACM, New York, NY, USA (2016)
19. Orlandi, F., Breslin, J., Passant, A.: Aggregated, interoperable and multi-domain user profiles for the social web. In: Proceedings of the 8th International Conference on Semantic Systems. pp. 41–48. ACM (2012)
20. Piao, G., Breslin, J.G.: Analyzing Aggregated Semantics-enabled User Modeling on Google+ and Twitter for Personalized Link Recommendations. In: User Modeling, Adaptation, and Personalization. ACM (2016)
21. Piao, G., Breslin, J.G.: Exploring Dynamics and Semantics of User Interests for User Modeling on Twitter for Link Recommendations. In: 12th International Conference on Semantic Systems. ACM (2016)
22. Staab, S., Studer, R.: Handbook on Ontologies. Springer Publishing Company, Incorporated, 2nd edn. (2009)
23. Zarrinkalam, F., Kahani, M.: Semantics-enabled User Interest Detection from Twitter. In: 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (2015)