User Modeling on Twitter with WordNet Synsets and DBpedia Concepts for Personalized 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 of individual users on the Social Web platforms such as Twitter plays a significant role in providing personalized recommendations and filtering interesting information from social streams. Recently, researchers proposed the use of concepts (e.g., DBpedia entities) for representing user interests instead of word-based approaches, since Knowledge Bases such as DBpedia provide cross-domain background knowledge about concepts, and thus can be used for extending user interest profiles. Even so, not all concepts can be covered by a Knowledge Base, especially in the case of microblogging platforms such as Twitter where new concepts/topics emerge everyday.

In this short paper, instead of using concepts alone, we propose using *synsets* from WordNet and *concepts* from DBpedia for representing user interests. We evaluate our proposed user modeling strategies by comparing them with other bag-of-concepts approaches. The results show that using synsets and concepts together for representing user interests improves the quality of user modeling significantly in the context of link recommendations on Twitter.

Keywords

User Modeling; Personalization; User Interest Profiles

1. INTRODUCTION

User modeling on Social Web platforms such as Twitter, which aims at building user interest profiles, has been proven to be an important way for capturing user interests from User-Generated Content (UGC). The generated user interest profiles then can be used for dealing with the *information overload* problem for providing personalized recommendations. Defining a way of representing user interests is an essential step in building user interest profiles. Previous work either used *bag-of-words*, *topic modeling* or *bag-of-concepts* approach to represent user interests. Bag-of-concepts approach uses concepts for representing user interests. For example, given the first sample tweet posted by a user named `Bob` in Table 1, we know that the user is interested in *entities* such as `dbpedia`[1]`:The_Black_Keys` and `dbpedia:The_Wombats`. The *bag-of-concepts* approach has been preferred in recent studies [10] so as to exploit background knowledge of concepts from a Knowledge Base (KB) (defined as the combination of an ontology and instances of the classes in the ontology) for extending user interests. For instance, based on the background knowledge from DBpedia, we can further infer that `Bob` is interested in `dbpedia:Indie_rock` as both `dbpedia:The_Wombats` and `dbpedia:The_Black_Keys` are pointing to `dbpedia:Indie_rock` via the property `dbpedia-owl`[2]`:genre`. In what follows, by a *concept* we mean an *entity* or *category* from a KB (e.g., DBpedia[3]) for representing user interests.

Table 1: Two sample tweets posted by Bob

#1	*My Top 3 #lastfm Artists: Eagles of Death Metal(14), The Black Keys(6) & The Wombats(6)*
#2	*Just completed a 3.89 km ride. We're gonna need more...*

Although Knowledge Bases such as DBpedia provide rich semantics from background knowledge for representing and propagating user interests, they cannot cover all existing and emerging concepts. In addition, Knowledge Bases lack full coverage for the lexicographic senses of lemmas, which can be provided by WordNet instead. For example, in the case of the second tweet posted by `Bob`, we cannot extract any concept from the tweet. To circumvent this drawback, we propose using WordNet[4] synsets and DBpedia concepts together for representing user interests. Synsets in WordNet are unordered sets of synonyms - words that denote the same concept and are interchangeable in many contexts. By doing so, from the second tweet, we can extract synsets such as: s_1=[*kilometer, kilometre, km, klick (a metric unit of length equal to 1000 meters (or 0.621371 miles))*] and s_2 = [*drive, ride (a journey in a vehicle (usually an automobile))*], which denote the user interests that would be missed if a concepts-alone approach was used.

Therefore, the goal of the study is to investigate the effectiveness of using WordNet synsets and DBpedia concepts

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

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

[3]http://wiki.dbpedia.org

[4]https://wordnet.princeton.edu

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

together for representing user interests. Proposed user modeling strategies are evaluated in the context of link recommendations on Twitter, by comparing our strategies and bag-of-concepts approaches using concepts alone for user modeling.

2. RELATED WORK

A line of work has been proposed to use *concept-based* representations of user interests using a KB from Linked Data (e.g., Freebase, DBpedia) [2, 8, 10]. This line of work goes beyond other approaches, such as *bag-of-words* [7] and *topic modeling* [4] which focus on words and cannot provide semantic information and relationships among these words. Abel et al. [1] showed that *entity-based* user interest profiles perform better than other user modeling strategies such as *hashtag-* and *topic-based* user profiles on Twitter in the context of news recommendations. Orlandi et al. [8] investigated *entity-based* user profiles and *category-based* user profiles based on the category information of entities from DBpedia. Besides 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. The results based on a user study showed that *category-based* user profiles have similar performance to the *entity-based* ones and both of them outperform the *word-based* approach as a baseline. On top of that, Piao et al. [10] proposed a mixed approach using *entity-* and *category-based* user profiles and presented its effectiveness compared to other user modeling strategies. Our work differs in that we represent user interests using WordNet synsets and DBpedia concepts together instead of using concepts alone. However, as we are using synsets in addition to concepts for representing user interests, interest propagation methods that were developed for concept-based user interests profiles can be applied to our interest profiles in the same way.

3. CONTENT-BASED USER MODELING

Interest Representation. In this work, we use synsets from WordNet and concepts from DBpedia for representing the interests of users. The generic model for interest profiles representing users is specified in Definition 1.

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

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

Here, $I = \{s_1, \ldots, s_k\} \cup \{c_1, \ldots, c_m\} = \{i_1, \ldots, i_n\}$ denotes the set of synsets in WordNet and concepts in DBpedia, and U denotes users.

The process of building user interest profiles using our user modeling framework is presented in Figure 1. It consists of three main steps as follows.

Interest Extraction. As we use WordNet synsets and DBpedia concepts for representing user interests, the first step is to extract synsets and concepts from UGC. In the same way from other bag-of-concept approaches, the concepts are extracted using NLP APIs such as the Aylien API[5]. To

[5]http://aylien.com/

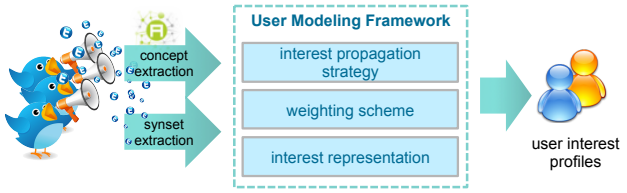


Figure 1: The process of generating user interest profiles

Algorithm 1: The WordNet-based WSD algorithm for tweets

input : a polysemous word w in a tweet t
output: the proper synset of w

1 $C \leftarrow \{w_1, \ldots, w_n\}$; // C is the context of w, i.e., other words in t with w
2 $X \leftarrow \{s_1, \ldots, s_k\}$; // X is the set of candidate synsets for w returned by WordNet
3 $s \leftarrow null$; // s is the synset to be returned
4 $score \leftarrow 0$; // $score$ is the similarity score assigned to s regarding the context C
5 $T \leftarrow \emptyset$; // T is the set of all candidate synsets for all words in C
6 **for** $w_j \in C$ **do**
7 **if** $POS(w_j) = POS(w)$ **then**
8 $X_j \leftarrow \{s_{j1}, \ldots, s_{jm}\}$;
9 $T \leftarrow T \cup X_j$;
10 **for** $s_i \in X$ **do**
11 **for** $s_h \in T$ **do**
12 $score_{ih} \leftarrow SINSIM(s_i, s_h)$; // computing similarity scores between s_i and every synset $s_h \in T$
13 **if** $score_{ih} \geq score$ **then**
14 $score \leftarrow score_{ih}$;
15 $s \leftarrow s_i$; // s is the synset $s_i \in X$ having the highest similarity score regarding the synsets T
16 **return** s

extract WordNet synsets, a WordNet-based Word Sense Disambiguation (WSD) algorithm [3] (see Algorithm 1), which was developed in the context of movie recommendations, has been adapted. This method extracts the WordNet synset for a word in terms of a context. In our scenario, the context of a word w is the set of words appearing in the same tweet with w (line 1) and having the same Part-Of-Speech (POS) as w. For a given tweet, our user modeling framework preprocesses with tokenization, POS tagging and lemmatization, and then uses Algorithm 1 for extracting all synsets for words based on their context. The similarity between any two synsets in Algorithm 1 (line 12) is measured as follows [5]:

$$SIMSIM(s_a, s_b) = -\log(N_p/2D) \quad (2)$$

where N_p is the number of nodes in the shortest path p from s_a and s_b, and D is the maximum depth of the taxonomy.

Weighting Scheme. We use the Term Frequency - Inverse Document Frequency (TF-IDF) as the weighting scheme for synsets and concepts: $w(u,i) = f(u,i) \times \log \frac{M}{m_i}$, where

$f(u,i)$ denotes the frequency of a synset or concept i in a user's tweets, M is the total number of users, and m_i is the number of users interested in i (i.e., who mentioned i in their tweets). Finally, we further normalize user interest profiles so that the sum of all weights in a profile is equal to 1: $\sum_{i \in I} w(u,i) = 1$.

Interest Propagation. As we are using both WordNet synsets and DBpedia concepts for user modeling, we can exploit background knowledge from DBpedia with respect to the extracted concepts for propagating user interests using existing methods. We use an extension strategy using category information from DBpedia for concepts with a discounting strategy for the extended interests as follows [10]:

$$CategoryDiscount = \frac{1}{\alpha} \times \frac{1}{\log(SP)} \times \frac{1}{\log(SC)} \quad (3)$$

where: *SP = Set of Pages belonging to the Category, SC = Set of Sub-Categories*. *SP* and *SC* discount the category in the context of DBpedia. Thus, an extended category is discounted more heavily if it is a general one (i.e., the category has a great number of pages or sub-categories). In addition, α is a parameter for discounting the extension from original user interest profiles (we set $\alpha = 2$ as in [10]).

4. EXPERIMENTAL EVALUATION

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 as in existing literature, we adopt a lightweight content-based algorithm where the recommendation algorithm recommends links according to their *cosine* similarity with a given user profile [1, 10].

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.

Dataset. The dataset used in this experiment is from [10]. The Twitter dataset includes all tweets published by 480 *active* users on Twitter (a user is *active* if the user published at least 100 posts [6,10]). The main details of the dataset are presented in Table 2. 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[6]). 322 out of 480 users met the criteria who published 247,676 tweets in total.

4.1 Evaluation methodology

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. Tweets older than two weeks (i.e., excluding those from the last two weeks, as used for ground truth), were used for constructing user profiles.

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

Table 2: 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

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 [1, 2, 8].

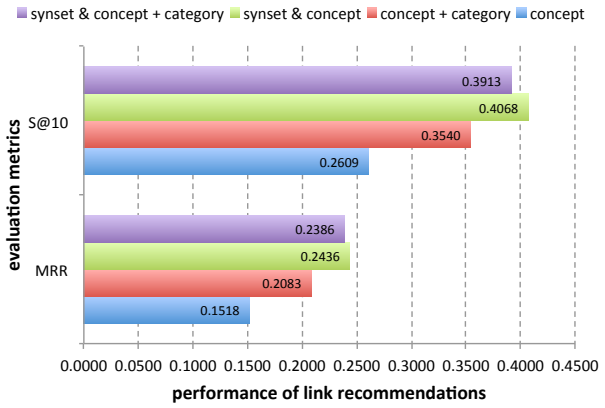
- **MRR** The *MRR* (Mean Reciprocal Rank) indicates at which rank the first item *relevant* to the user occurs on average.
- **S@N** The Success at rank N (S@N) stands for the mean probability that a relevant item occurs within the top-N of the ranking.
- **R@N** The Recall at rank N (R@N) represents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

We will focus on the top-10 list of recommendations for success, recall and precision as our recommendation system will list 10 link recommendations to a user. The *bootstrapped paired t-test*, which is an alternative to the paired t-test when the assumption of normality of the method is in doubt, is used for testing the significance where the significance level was set to 0.05 unless otherwise noted.

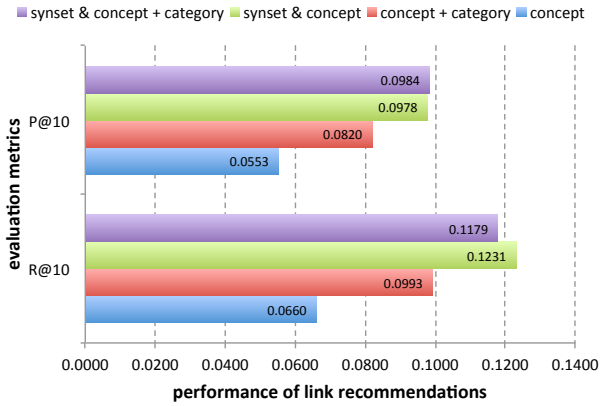
4.2 Results

To evaluate whether our new synset & concept-based user interest profiles outperform concept-based profiles, we use the concept-based user interest profiles (`P(concept)`) [1] and extended `P(concept)` using background knowledge from DBpedia (`P(concept+category)`) [10] as two baselines. Our approach is represented as `P(synset&concept)`, which uses synset and concepts for representing user interests. In addition, the synset & concept-based user interest profiles extended with background knowledge are denoted as `P(synset&concept+category)`.

The results of link recommendations based on different user modeling strategies in terms of aforementioned four different evaluation metrics are presented in Figure 2. As we can see from the figure, there is a significant improvement for `P(synset&concept)` and `P(synset&concept+category)` compared to the concept-based approaches (`P(concept)` and `P(concept+category)`, $p < 0.05$). For example, the quality of recommendations is improved by `P(synset&concept)`



(a) MRR and success rate



(b) Precision and recall

Figure 2: Performance of link recommendations based on different user modeling strategies

56% and 61% in terms of S@10 and MRR, 77% and 87% in terms of P@10 and R@10, compared to using `P(concept)`. Similarly, using `P(synset&concept+category)` improves the recommendation performance 11% and 15% in terms of S@10 and MRR, 20% and 19% in terms of P@10 and R@10 compared to using `P(concept+category)`. This indicates that using synsets and concepts together is beneficial for user modeling on Twitter instead of using concepts alone.

It is also interesting to observe that `P(synset&concept)`, which uses synsets and concepts together without any interest propagation, has competitive performance compared to the one using the same interest representation and propagating interests with background knowledge (`P(synset&concept+category)`).

5. CONCLUSIONS

In this paper, we were interested in examining if using synsets from WordNet and concepts from DBpedia for representing user interests can improve the quality of user modeling on Twitter. We presented our user modeling strategies using both synsets and concepts and evaluated them by comparing to other concept-based strategies in the context of link recommendations on Twitter. The study results showed that using synsets and concepts together for representing user interests improves the quality of user modeling significantly. In addition, user profiles with a rich interest representation without any interest propagation (`P(synset&concept)`) can provide competitive performance compared to `P(synset&concept+category)`. Future studies will focus on investigating user modeling strategies considering other dimensions (e.g., temporal dynamics of user interests) together [9].

6. 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).

7. REFERENCES

[1] 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.

[2] 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.

[3] M. Degemmis, P. Lops, and G. Semeraro. A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation. *User Modeling and User-Adapted Interaction*, 17(3):217–255, 2007.

[4] M. Harvey, F. Crestani, and M. J. Carman. Building User Profiles from Topic Models for Personalised Search. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2309–2314, 2013.

[5] C. Leacock and M. Chodorow. Combining Local Context and WordNet Similarity for Word Sense Identification. *An Electronic Lexical Database*, pages 265–283, 1998.

[6] 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.

[7] 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.

[8] 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.

[9] G. Piao. Towards Comprehensive User Modeling on the Social Web for Personalized Link Recommendations. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, UMAP '16, pages 333–336, New York, NY, USA, 2016. ACM.

[10] 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*, pages 105–109. ACM, 2016.