*This paper has been accepted as Doctoral Consortium paper in the UMAP 2016 conference

Towards Comprehensive User Modeling on the Social Web for Personalized Link Recommendations

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

ABSTRACT

User modeling for individual users on the Social Web plays a significant role and is a fundamental step for personalization as well as recommendations. Previous studies have proposed various user modeling strategies in different dimensions such as (1) *interest representation*, (2) *interest propagation*, (3) *content enrichment* and (4) *temporal dynamics of user interests*. This research mainly focuses on the first two dimensions *interest representation* and *propagation*. In addition, we also investigate the combination of these four dimensions and their synergistic effect on the quality of user modeling. Different user modeling strategies will then be evaluated in the context of personalized link recommender systems using standard evaluation methodologies such as Mean Reciprocal Rank (MRR), recall (R@N) and success (S@N) at rank N.

1. INTRODUCTION AND MOTIVATIONS

With the advent of Web 2.0, users participate more and more in the evolution of the Web and become big contributors to data on the Web. The term "Social Web" indicates a paradigm shift from a machine-centered view of the Web towards a more user- or community-centered view [1]. In this context, a better understanding of user interactions and then using this to extract, analyze and represent the information about users on the Web is crucial for systems providing personalized services as well as recommendations. One of the challenges for service providers is to provide accurate recommendations and personalization without having to explicitly ask for users' feedback or make users wait for reliable recommendations only after a long initial training period on the system (the so-called *cold start* problem). To overcome these challenges, it is important to create qualitative and quantitative user models on the Social Web.

Interest representation. The representation of user interests is a major dimension for user modeling on the Social Web. Previous work either used *bag-of-words*, *topic model* or *bag-of-concepts* approach to represent user interests. Bag-of-concepts approach uses concepts for representing user interests. For example, given a tweet posted by a user in Figure 1, we know that the user is interested in entities such as `dbpedia`[1]`:The_Wombats` and `dbpedia:The_Black_Keys`. Recently, studies have been using *bag-of-concepts* approach as background knowledge of concepts from a Knowledge Base (KB) (defined as the combination of an ontology and instances of the classes in the ontology [19]) can be leveraged for extending user interests. By a *concept* we mean an *entity*, *category* or *class* from a KB (e.g., DBpedia[2]) for representing user interests.

Figure 1: A sample tweet

Interest propagation with background knowledge. A Knowledge Base provides cross-domain background knowledge about concepts. For instance, DBpedia is the semantic representation of Wikipedia[3] and is a KB consisting of a large amount of facts about concepts. Compared to traditional taxonomies or lexical databases (e.g. WordNet [13]) it provides a larger set of concepts and the relationships among these concepts, continuously updated by the Wikipedia community. Therefore, DBpedia is helpful for extending user interest profiles to provide more semantic information about users. For instance, in the example in Figure 1, we can further infer that the user 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`[4]`:genre`.

Enrichment. One of the problems in user modeling on the Social Web is short posts (e.g., 140 characters for tweets) generated by users. To better understand the semantic meaning of short posts, enriching the posts with different sources (e.g., embedded links in the posts [4,10]) is important.

Dynamics. The interests of users might change over time. Previous studies showed that considering the time decay of user interests improves the quality of user modeling.

[1]The prefix `dbpedia` denotes http://dbpedia.org/resource/
[2]http://wiki.dbpedia.org
[3]https://en.wikipedia.org/wiki/Main_Page
[4]The prefix `dbpedia-owl` denotes http://dbpedia.org/ontology/

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

In this thesis, we propose using synsets from WordNet and concepts from DBpedia for representing user interests. In addition, we investigate different methods using background knowledge from WordNet as well as DBpedia for extending user interests. On top of that, we look at whether the quality of user modeling can be improved by considering temporal dynamics and enrichment using existing approaches. Our hypothesis is that the final extended user interest profiles, which are based on enriched user profiles considering temporal dynamics, can represent current user interests better and improve the quality of link recommendations.

2. RELATED WORK

Representation of user interests. A line of work has been proposed to use *concept-based* representation of user interests using a KB from Linked Data (e.g., Freebase, DBpedia) [4, 5, 15] or an encyclopedia such as Wikipedia [9, 11, 12]. This line of work goes beyond other approaches, such as *bag-of-words* [14] and *topic modeling* [8] which focus on words and cannot provide semantic information and relationships among these words. Thus, we start with *concept-based* user profiles in our work, and mainly focus on related work with respect to the *concept-based* user profiles here.

Semantics-enabled user modeling. On top of the *concept-based* user profiles, researchers have proposed using rich semantic information from a KB to extend the interests of users. Abel et al. [5] proposed using DBpedia to extend user profiles with respect to point of interests (POI) and showed that the extended POI profiles outperform the original user profiles without any extension in the context of POI recommendations. Orlandi et al. [15] proposed *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. Even so, *category-based* profiles revealed much more information about user interests and the authors claimed that the *category-based* profiles might be helpful in the context of recommender systems [15]. However, they did not further evaluate those user modeling strategies in the context of recommendations.

Enrichment for short messages. The length of User-Generated Content (UGC) on Online Social Networks (OSNs) such as Twitter[5] is usually short. For example, Twitter messages (tweets) are limited to 140 characters which makes it difficult to detect the semantics of these messages. Researchers [4, 10] have been exploring embedded links 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.

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, 15]. 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 using their entire historical UGC. Instead of extracting user profiles within a certain period, researchers have proposed using a 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. However, every work proposed its own evaluation method for its measure and there was no study provides a comparison of different methods proposed in the literature.

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

3. PROBLEM STATEMENT AND RESEARCH GOALS

While related work reveals several insights regarding the user modeling process on the Social Web, there are limitations with respect to the representation of user interests and the interest propagation using a KB such as DBpedia.

Interest representation. Although a KB provides a great amount of semantic information about concepts, not everything (which is a concept in KB) is covered by a KB. This is problematic especially in the case of the Social Web where new concepts/topics emerge everyday. In this regard, user interests might not be fully extracted and covered using concepts in a KB. In addition, Knowledge Bases such as DBpedia lack full coverage for the lexicographic senses of lemmas, which can be provided by a lexical database such as WordNet.

Interest propagation using DBpedia. Despite the focus on leveraging category information for entities from DBpedia in the literature, background knowledge of concepts in DBpedia is provided through different types of information. For example, category information in DBpedia relies on the category system of Wikipedia to capture the idea of a "theme", i.e., a subject of an entity while class information about entities uses classification from YAGO [20]. In addition, there are many semantically related entities which are connected via various properties. Thus, different types of information (e.g., *categories*, *classes* and *connected entities via different properties*) should be considered all together for extending user interest profiles.

Comprehensive user modeling. In addition, there is a need to study comprehensive user modeling strategies considering different dimensions together to investigate whether the quality of user interest profiles can be improved further due to any synergetic effect of the combination. For instance, user interests extracted through the embedded links within UGC can also further extended using background knowledge from DBpedia to reveal more information about users.

Research questions to address these research gaps are summarized as follows:

RQ 1*. (How) can we exploit different aspects (e.g., categories, classes, connected entities via different properties) of background knowledge from DBpedia for extending user interest profiles?*

RQ 2*. Does the use of the lexical database and knowledge base for the representation of user interests improve the quality of user modeling?*

RQ 3*. Is there any synergetic effect on user modeling by combining different dimensions? In other words, does the extended user interest profiles, which are based on enriched user profiles considering temporal dynamics, improve the quality of user modeling?*

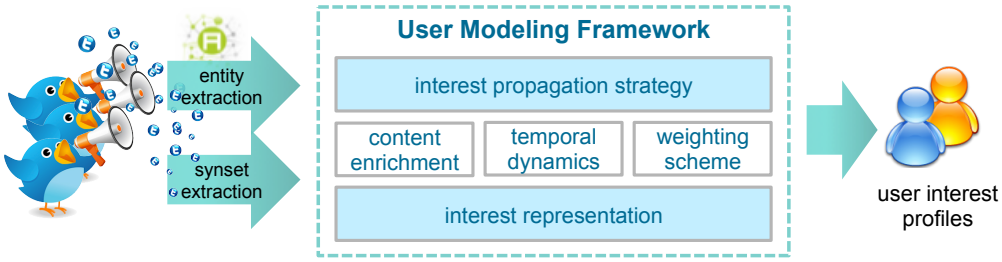

Figure 2: The process of generating user interest profiles using our user modeling framework

4. EVALUATION SETUP

Dataset. To evaluate different user modeling strategies, we need to collect a dataset of OSN users. Based on the `about.me`[6] dataset collected from our previous studies, we further extracted UGC of users from Twitter and Google+. `about.me` offers registered users a simple platform from which to link multiple online identities of them including popular OSNs such as Twitter, Google+[7] etc. We crawled a large number of OSN accounts of users from `about.me`, which consist of 247,630 public profile pages of users and their OSN account information. Furthermore, we randomly selected 480 users from `about.me` dataset who had been using Twitter and Google+, and collected all the UGC of those 480 users from these two OSNs for our study.

Evaluation Methodology. The quality of different user modeling strategies can be evaluated by investigating how these strategies affect the performance of an application which works on the basis of user interests. In the same way from previous studies [5], we evaluate different user modeling strategies in the context of a personalized recommender system where the recommended items are links. The ground truth and a set of candidate links are built by the links shared by users via their UGC. The recommender system then recommends links based on a recommendation algorithm with different user profiles generated by different user modeling strategies as input. The quality of the top-N recommendations is measured via standard evaluation metrics such as *MRR* (Mean Reciprocal Rank), success rate (S@N), precision and recall (P@N and R@N).

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 recommendations. We do not aim to optimize recommendation quality, but are interested in comparing the quality achieved by the same recommendation algorithm when inputing different types of user profiles. Therefore, we apply a lightweight content-based algorithm, like the one used in [5], that recommends items according to their *cosine* similarity with a given user profile.

5. APPROACH AND PROGRESS TO DATE

In this work, we use synsets from WordNet or concepts from DBpedia for representing the interests of users. Synsets in WordNet are unordered sets of synonyms - words that denote the same concept and are interchangeable in many contexts. The generic model for interest profiles representing users is specified in Definition 1.

[6]http://about.me
[7]https://plus.google.com

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.

Here, I denotes the set of synsets in WordNet and concepts in DBpedia, and U denotes users, respectively.

The general process of generating user interest profiles is presented in Figure 2. Entities in UGC are extracted using the Aylien API[8] and synsets can be extracted by adopting a previous method for generating WordNet-based user profiles [7]. On top of the representation of user interest profiles, our user modeling framework is able to adhere to different strategies with respect to a specific interest propagation strategy as well as other existing strategies regarding enrichment and dynamics of user interests.

Progress to date. Until now, we have focused on extending strategies using DBpedia for concept-based user profiles. Preliminary results of the extending strategies were published in [16]. In that work, we evaluated two *category-based* user modeling strategies with DBpedia from [15]; one is *category-based* user profiles *replacing* the corresponding *entity-based* user profiles, and the other one is *category-based* user profiles with a discounting strategy for the weight of each category based on its specificity within DBpedia graph. In addition, we investigated a mixed approach that combines the *entity-* and *category-based* user profiles. Results showed that the combined user profiles with the discounting strategy for extended categories provide the best performance in terms of link recommendations. In contrast, both *category-based* user modeling strategies with/without a discounting strategy did not outperform *entity-based* user profiles.

In terms of $w(u, c)$ in Definition 1, previous studies have been applying $CF_u(c)$ (frequency of c in u's UGC, e.g., tweets) as the weighting scheme $w(u, c)$. However, our recent experiment using $CF_u(c) \times IDF_u(c)$ as the weighting scheme showed significant improvement over using $CF_u(c)$ where $IDF_u(c) = \log[\#\ of\ all\ users]/[\#\ of\ users\ interested\ in\ c]$.

6. FUTURE PLAN AND CONTRIBUTIONS

Although we leveraged DBpedia categories for user modeling with a discounting strategy, there is much more information about entities in DBpedia via various properties. As a discounting strategy for extended entities from DBpedia plays a significant role, we need to derive a discounting strategy for extending user profiles with various properties in DBpedia. For instance, we can apply discounting strategies based on the lessons learned from similarity-based Linked

[8]http://aylien.com

Data-enabled recommender systems [17] as the normalizations of properties in this research area play the same role as ours. As a further step, we plan to adopt the spreading activation theory [18] for interest propagation by leveraging different types of information from DBpedia.

In the near future, we would investigate and evaluate the proposed interest representation model, which is using WordNet synsets and DBpedia concepts together for representing user interests, in the context of link recommendations using our collected Twitter and Google+ dataset.

Finally, we aim to study combined user modeling strategies considering different dimensions mentioned in the related work to derive more sophisticated user interest profiles in the context of personalized link recommendations.

The main contributions of this research can be summarized as follows.

- We propose a method exploiting different types of information from DBpedia for extending user profiles.
- We investigate and evaluate the representation of user interests by using synsets from WordNet and concepts from DBpedia for user modeling.
- Importantly, this research study will provide a comprehensive user modeling framework combining aforementioned different dimensions to retrieve extended user interest profiles, which are based on enriched user profiles considering temporal dynamics.

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). In addition, I would like to thank my supervisor Dr. John G. Breslin for his support and feedback.

8. REFERENCES

[1] F. Abel. *Contextualization, User Modeling and Personalization in the Social Web–From Social Tagging via Context to Cross-System User Modeling and Personalization*. PhD thesis, Leibniz University of Hanover, 2011.

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

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] G. A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] G. Salton and C. Buckley. On the use of spreading activation methods in automatic information. In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '88*, pages 147–160, New York, New York, USA, 1988. ACM Press.

[19] S. Staab and R. Studer. *Handbook on Ontologies*. Springer Publishing Company, Incorporated, 2nd edition, 2009.

[20] 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.