Analyzing Aggregated Semantics-enabled User Modeling on Google+ and Twitter for Personalized 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

In this paper, we study if reusing Google+ profiles can provide reliable recommendations on Twitter to resolve the *cold start* problem. Next, we investigate the impact of giving different weights for aggregating user profiles from two OSNs and present that giving a higher weight to the targeted OSN profile for aggregation allows the best performance in the context of a personalized link recommender system. Finally, we propose a user modeling strategy which combines *entity-* and *category-based* user profiles using with a discounting strategy. Results show that our proposed strategy improves the quality of user modeling significantly compared to the baseline method.

1. INTRODUCTION

With the growing popularity of Online Social Networks (OSNs) and the increased number of OSNs that users tend to use, studies of reusing or aggregating different OSN profiles for user modeling and then using it for recommendations have been widely conducted. Abel et al. [4] described *tag-based* user profiles from OSNs such as Delicious[1], StumbleUpon[2] and Flickr[3] and used the profile of other services for recommendations in the targeted OSN (e.g., reusing user profiles from Delicious for recommendations on Flickr in a cold start situation). The *targeted OSN* denotes the OSN where we recommend items to the users. Different OSNs have different characteristics. The social bookmarking and photo sharing OSNs in the study [4] have a great amount of tags in addition to their main content, and therefore *tag-based* profiles have been used. However, in a microblogging service like Twitter or other general OSNs like Google+[4], the main content usually consists of short messages, and therefore *entity-based* user profiles (i.e., user interests are represented by entities, e.g., `dbpedia`[5]`:Steve_Jobs`, `dbpedia:Apple_Inc.`) have been used [1]. On top of entity-based user profiles, researchers [3,15] proposed extending user profiles with background knowledge from Linked Data [5] (e.g., DBpedia [10]) since it provides rich semantic information about entities. We propose a mixed approach using *entity-* and *category-based* user profiles and evaluate the user modeling strategy in the context of link recommendations (category-based user profiles represent user interests using categories, e.g., `dbpedia:Category:Electronics_companies` for the entity `dbpedia:Apple_Inc.`).

The main contributions of our work are as follows: (i) an investigation of the benefits of reusing Google+ profiles for personalized link recommendations on Twitter (Section 5.1), (ii) a study of aggregation strategies with different weighting scheme for different OSN profiles (Section 5.2), and (iii) the evaluation of our mixed approach for extending user profiles using background knowledge from DBpedia (Section 5.3).

2. RELATED WORK

Mehta et al. [13] proposed cross-system personalization approaches, which aim to make recommender systems more robust against spam and cold start problems. However, they could not evaluate their methods on the Social Web data. In [4], the authors investigated aggregated *tag-based* profiles from Delicious, StumpbleUpon and Flickr. They investigated how the aggregated *tag-based* user profiles impact on tag and resource recommendations, especially in cold start situations. Different user modeling strategies should be applied to different types of OSNs.. The same authors from [4] proved that the *entity-based* user profiles outperform other approaches such as *hashtag-* or *topic-based* user profiles on Twitter [1]. However, they did not evaluate aggregated *entity-based* profiles from general OSNs such as Google+ or Twitter further. In this regard, it has not been shown if the cold start problem in a more general OSN can be resolved by the aggregated entity-based profiles (as has been shown for tag-based profiles on social tagging systems).

To aggregate user profiles from different OSNs, previous studies applied the same weight to each OSN profile [4,15] but did not look at different weights for aggregating OSN profiles. A recent survey [6] also pointed out that the method of giving equal weights to different OSNs for aggregating

[1] https://www.delicious.com
[2] https://www.stumbleupon.com
[3] https://www.flickr.com
[4] https://plus.google.com

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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

UMAP '16, July 13-17, 2016, Halifax, NS, Canada
© 2016 ACM. ISBN 978-1-4503-4370-1/16/07. . . $15.00
DOI: http://dx.doi.org/10.1145/2930238.2930278

profiles might not be enough, and a sophisticated model could be derived based on the specific needs for recommendations.

In the past years, user modeling strategies leveraging background knowledge (e.g., DBpedia) for extending user profiles have been developed [3, 9, 14, 15]. Abel et al. [3] proposed using Linked Data to extend user profiles and proved that extending user profiles with rich information from Linked Data can improve user modeling in terms of point of interest (POI) recommendations. Orlandi et al. [15] proposed *category-based* user profiles based on category information for entities from DBpedia. Besides a straightforward extension that gives equal weight to each extended category with respect to an entity [3], they also proposed a discounting strategy for those extended categories. Although *category-based* and *entity-based* user profiles showed similar performance in their user study, the authors [15] claimed that *category-based* user profiles produced almost seven times more user interests and might be helpful in the context of recommender systems. However, they did not further evaluate those user modeling strategies in the context of recommendations and left it as future work. Our work is more similar to [3, 15] in terms of the knowledge base that has been exploited.

3. CONTENT-BASED USER MODELING

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

Definition 1. The profile of a user $u \in U$ is a set of weighted DBpedia entities where with respect to the given user u for an entity $e \in E$ its weight $w(u, e)$ is computed by a certain function w.

$$P_u = \big\{\big(e, w\big(u, e\big)\big) \mid e \in E, u \in U\big\} \tag{1}$$

Here, E and U denote the set of entities in DBpedia and users respectively. We apply occurrence frequency as the weighting scheme $w(u,\ e)$, which means that the weight of an entity (interest) is determined by the number of OSN activities in which user u refers to the entity e. 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`. We further normalize user profiles so that the sum of all weights in a profile is equal to 1: $\sum_{e_i \in E} w(u, e_i) = 1$.

To get aforementioned user profiles, we implemented a user modeling framework that retrieves user profiles from User-Generated Content (UGC). Our framework features three main components:

Link Extractor. Given User-Generated Content (tweets and Google+ posts in this study), the component extracts all links (URLs) in the content by using a defined regex pattern. Furthermore, this component expands a URL from a shortened form (e.g., `http://t.co/Is6l9ODiny`), which is a common practice in OSNs.

Entity Extractor. This component extracts all DBpedia entities within UGC using the `Aylien API`[6]. In addition, it is used for retrieving DBpedia entities in the content of the links which were extracted by the `Link Extractor`.

Profile Generator. Based on the extracted entities, our framework provides a method for generating user profiles that might adhere to aggregating strategies with different weights for different OSN profiles as well as extending strategies with background knowledge from DBpedia.

4. DATASET

Users tend to have multiple social identities in different OSNs [11]. To retrieve the ground truth data (i.e., users who are using both Google+ and Twitter), we obtained OSN accounts of users from `about.me`[7]. We crawled 247,630 public profile pages from `about.me` during December 2014 that have at least two external links. In our dataset, the number of different OSNs (29) and the average number that each person participates in (4.48) are both larger than the numbers from a previous study [11], which are 15 and 3.92 respectively.

As `about.me` dataset only contains OSN accounts of users, we need to retrieve all UGC from selected OSNs for our study. We chose Google+ and Twitter for our study due to (1) their higher degrees of openness, and (2) UGC from OSNs such as tweets has been demonstrated to be a good indicator for determining user interests in [2, 15]. As we were interested in analyzing aggregated user profiles from Google+ and Twitter, we randomly selected 480 *active* users from the `about.me` dataset who had been using both OSNs. We then extracted their UGC as well as all links shared with those UGC using the aforementioned framework. Similar to other studies [8, 12], we define that a user is *active* if the user published at least 100 posts (i.e., tweets and Google+ posts). In addition, we selected users who shared at least 10 links via their tweets to construct ground truth links. After all, there were 429 *active* users in the dataset for the experiment (41 users did not have 10 links in their recent posts). The dataset is available via the supporting website of this paper [16].

5. USER MODELING FOR PERSONALIZED LINK RECOMMENDATIONS

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 this regard, we adopt a lightweight content-based algorithm as the recommendation algorithm that recommends links according to their *cosine* similarity with a given user profile.

Definition 2. Recommendation Algorithm: given P_u and a set of candidate links $\mathrm{N} = \big\{P_{i1}, ..., P_{in}\big\}$, 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 the 10 latest links shared via the user's tweets. We used 10 links of each user from 429 users, as well as the links shared by other users but not shared by 429 users in the dataset, for constructing candidate links. As a result, the set of candidate links consists of 5,165 distinct links. The rest of tweets and Google+ posts before the recommendation time were all used for constructing user profiles. The quality of the top-N recommendations was measured via the Mean Reciprocal Rank (*MRR*) and the

[6] http://aylien.com

[7] https://about.me

Recall at rank N (R@N), which have been widely used in the literature [3, 15]. *MRR* indicates at which rank the first item *relevant* to the user occurs on average, and R@N represents the mean probability that *relevant* items are retrieved within the top-N recommendations. We used the *bootstrapped paired t-test* for testing the significance where the significance level was set to 0.05 unless otherwise noted.

5.1 Using Google+ profiles for recommendations on Twitter in a cold start situation

RQ1: Can we reuse entity-based user profiles from Google+ to recommend links on Twitter? In [4], the authors used *tag-based* user profiles from other OSNs such as Delicious to recommend items on Flickr and showed that using other social bookmarking OSN profiles can improve recommendations in the targeted photo sharing OSN in cold start situations. In the same way, we used Google+ profiles of users for recommendations on Twitter especially in a cold start situation. To answer the research question (*RQ1*), we blinded out Twitter profiles of users and used only Google+ profiles of them to provide link recommendations on Twitter. We used the top-popular item recommender (`TopPop`) as a baseline, which is a common practice for a user in cold start until the user has interacted with the service enough [7].

Results. Figure 1 shows the results with respect to different evaluation methods for link recommendations by using Google+ profiles (`Gonly`) with the recommendation algorithm (Definition 2) and the results with `TopPop` recommendations on Twitter. As we can see from the figure, `Gonly` outperforms the baseline method `TopPop` significantly in terms of all evaluation methods. The value of *MRR* is 22.91%, which indicates that by using Google+ profiles, users can find a preferred link in the top 5 recommendations on Twitter on average. The results show that using Google+ profiles improves the quality of link recommendations significantly ($p < 0.01$) compared to the baseline method, and achieves comparable performance to using Twitter profiles (`Tonly`). In line with the results from the study [4], our results show that we can reuse Google+ profiles of users to provide personalized recommendations on Twitter in the cold start situation.

5.2 Aggregated user modeling with different weighting strategies

RQ2: Do aggregated profiles giving a higher weight to the targeted OSN perform better than those with the same weight for each OSN? To study the impact of aggregating profiles with different weights on user modeling, we assessed different weights for Twitter profile (which is the targeted OSN here), between 1 and 10 in steps of 1 and compared to the user profile of Twitter without aggregation (`Tonly`) as a baseline. Where previous works improved over `Tonly` by giving equal weight to each OSN in the aggregated profile, our hypothesis is that we can improve this further by giving a higher weight to the targeted OSN profile.

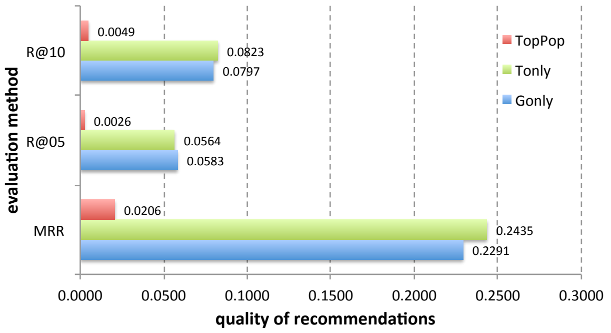
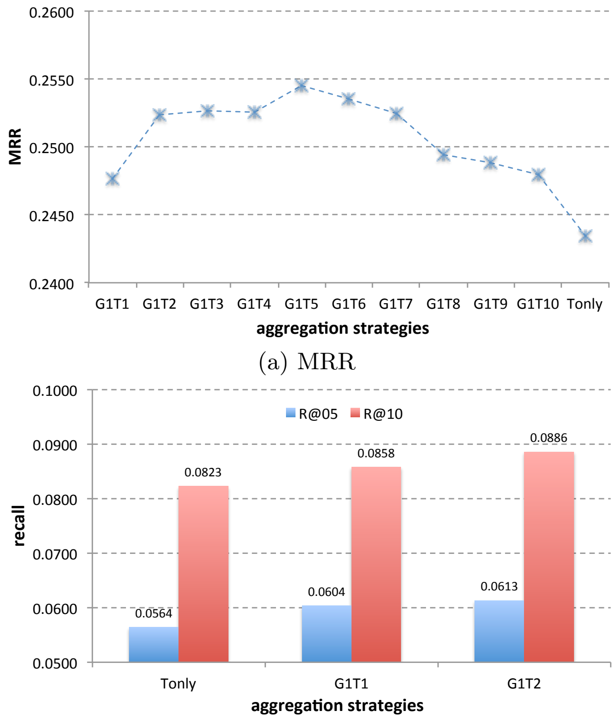


Figure 1: Performance of link recommendations on Twitter using Google+ profiles and the TopPop recommender



(a) MRR

(b) Recall

Figure 2: Performance of link recommendations based on different aggregating strategies for Google+ and Twitter

Results. Figure 2 (a) shows the performance of recommendations based on different weighting strategies in terms of *MRR*. `GmTn` denotes the weights `m` and `n` for Google+ and Twitter profiles respectively. For instance, `G1T1` denotes the aggregated profile with the same weight for Google+ and Twitter profiles while `G1T2` denotes the aggregated profile giving weight 1 for Google+ profile and weight 2 for Twitter profile. Finally, `Tonly` denotes the Twitter profile without any aggregation. As we can see from the Figure 2 (a), the performance of link recommendations begins to increase by giving a higher weight to the targeted OSN and then decreases if the weight is too high. Overall, `G1T5` performs best in terms of *MRR* and improves `Tonly` significantly while `G1T1` does not. Regarding the recall of recommendations (see Figure 2 (b)), `G1T2`, which gives a higher weight to the targeted OSN profile, performs best as well. Similar to *MRR* result, `G1T2` outperforms `Tonly` significantly in terms of both R@5 and R@10 while `G1T1` does not. While the weight for the targeted OSN profile is different for achieving the best performance in terms of different evaluation methods, the aggregated profile with a higher weight for the targeted OSN always performs best (i.e., `G1T5` and `G1T2` for *MRR* and recall respectively), and improves Tonly significantly. This indicates that aggregated profiles with a higher weight for the targeted OSN are required to achieve the best performance in terms of link recommendations.

Therefore, we conclude that the aggregated user profile giving a higher weight to the targeted OSN performs better than that of giving equal weight to each OSN and improves Tonly significantly.

5.3 Extended user modeling with categories from DBpedia

In this section, we evaluate two *category-based* user profiles from [15] compared to Tonly. In addition, we propose combined user profiles of *entity-* and *category-based* profiles (Tonly+T(x)) and evaluate them in the context of link recommendations. The two *category-based* user profiles from [15] and the combined profiles are as below:

T(Cat) [15]: A straightforward way of *replacing* Tonly with the categories from DBpedia applying the same weights of the corresponding entities in the *entity-based* profiles.

T(CatDiscount) [15]: Instead of the straightforward extension, this method applies a discounting strategy (Equation 2) for the extended categories from DBpedia.

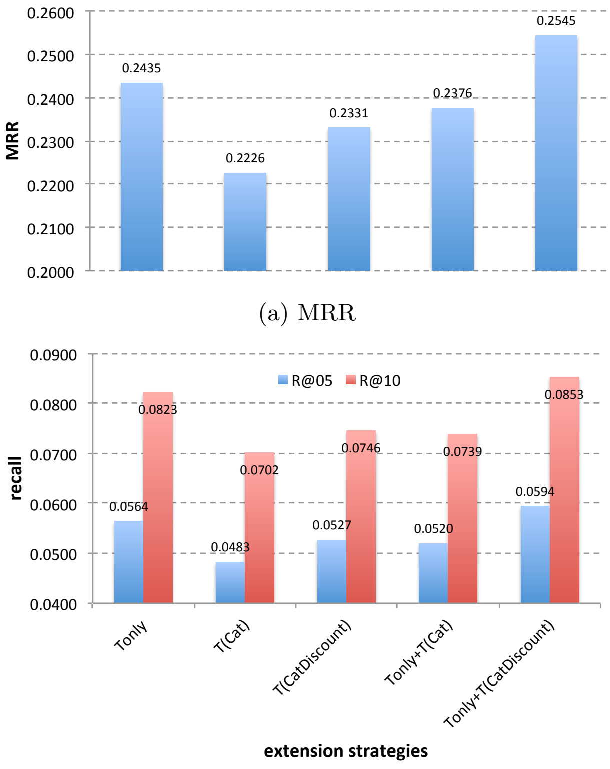
Tonly+T(x): This strategy combines the *entity-based* profiles (i.e., Tonly) as well as one of the *category-based* profiles mentioned above.

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

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, we add the parameter α which denotes the discount of the extended *category-based* user profiles for combining the *entity-based* and *category-based* user profiles. Thus, this parameter only has an effect on the combined user modeling strategies with the discounting strategy for the extended categories, i.e., **Tonly+T(CatDiscount)**. We set $\alpha = 2$ for this experiment.

Results. Figure 3 illustrates the recommendation performance of using different user modeling strategies based on category information from DBpedia as well as the performance of using Tonly in terms of *MRR* and recall. As depicted in Figure 3, Tonly+T(CatDiscount) achieves the best performance in the context of link recommendations and significantly outperforms Tonly in terms of all evaluation methods. In contrast, other strategies do not perform as well as Tonly. For instance, *category-based* user profiles (T(Cat) and T(CatDiscount)) and the combined user profiles with the straightforward extension of categories (Tonly+T(Cat)) do not outperform Tonly but decrease the performance of link recommendations.

Different from the hypothesis from [15], *category-based* user profiles do not perform better than *entity-based* user profiles in the context of recommender systems. However, the results show that the combined user profiles of *entity-* and *category-based* profiles with the discounting strategy (Tonly+T(CatDiscount)), improve the *entity-based* user profiles significantly and allow the best performance in terms of link recommendations compared to other user modeling strategies.



(a) MRR

(b) Recall

Figure 3: Performance of link recommendations based on extended user profiles using background knowledge from DBpedia

6. CONCLUSIONS AND FUTURE WORK

In this paper, we explored two dimensions of user modeling: (1) aggregating strategy of user profiles from different OSNs, and (2) extending strategy using background knowledge from DBpedia, and evaluated different strategies in the context of link recommendations. We investigated and proved the benefits of reusing Google+ profiles for link recommendations on Twitter in a cold start situation (refer to *RQ1*). Next, we studied different weighting strategies for aggregating Twitter and Google+ profiles. Unlike the approach from previous studies, results show that a higher weight must be given to the targeted OSN when aggregating profiles from different OSNs in order to achieve the best performance (refer to *RQ2*). Finally, we evaluated our mixed approach using entity- and category-based user profiles. Results show that our proposed user modeling strategy performs better than *category-based* user profiles as well as that with the straightforward extension strategy, and improves Tonly significantly. In the near future, we plan to investigate different aspects of DBpedia, such as classes and entities connected via different properties for user modeling.

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] 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, 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.

[3] 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.

[4] F. Abel, E. Herder, G.-J. Houben, N. Henze, and D. Krause. Cross-system user modeling and personalization on the social web. *User Modeling and User-Adapted Interaction*, 23(2-3):169–209, 2013.

[5] 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.

[6] K. Bontcheva and D. Rout. Making sense of social media streams through semantics: A survey. *Semantic Web*, 5(5):373–403, 2014.

[7] F. Cena, S. Likavec, and F. Osborne. Property-based interest propagation in ontology-based user model. In *User Modeling, Adaptation, and Personalization*, pages 38–50. Springer, 2012.

[8] 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, Rio de Janeiro, Brazil, 2013. International World Wide Web Conferences Steering Committee.

[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] 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.

[11] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon. What's in a name?: an unsupervised approach to link users across communities. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 495–504. ACM, 2013.

[12] 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.

[13] B. Mehta. *Cross System Personalization: Enabling personalization across multiple systems*. PhD thesis, 2008.

[14] F. Narducci, C. Musto, G. Semeraro, P. Lops, and M. Gemmis. Leveraging Encyclopedic Knowledge for Transparent and Serendipitous User Profiles. pages 350–352. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[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. Supporting website: details about datasets and additional findings, 2016.