Leveraging User Modeling on the Social Web with Linked Data

Fabian Abel, Claudia Hauff, Geert-Jan Houben, and Ke Tao

Web Information Systems, Delft University of Technology
{f.abel,c.hauff,g.j.p.houben,k.tao}@tudelft.nl

Abstract. Social Web applications such as Twitter and Flickr are widely used services that generate large volumes of usage data. The challenge of modeling the use and the users of such Social Web services based on their data has received a lot of attention in recent years. In this paper, we go a step further and investigate how the Linked Open Data (LOD) cloud can be leveraged as additional knowledge source in user modeling processes that exploit user data from the Social Web. Specifically, we introduce a user modeling framework that utilizes semantic background knowledge from LOD and evaluate it in the area of point of interest (POI) recommendations. For this purpose, we infer user preferences in POIs based on the users' behavior observed on Twitter and Flickr, combined with referable evidence from the Web of Data. We compare strategies that aggregate knowledge from two LOD sources: GeoNames and DBpedia. The evaluation validates the advantages of our approach; we show that the user modeling quality improves when LOD-based background information is included in the process.

1 Introduction

The Social Web is a gold mine for researchers and developers of user modeling techniques who investigate how user traces such as clicks, ratings, shared resources or textual contributions can be transformed into representations that are beneficial for a given application. For example, the status messages (so-called *tweets*) that people post on Twitter[1] can be exploited to feature personalized website recommendations or news recommendations [1,2]. To apply user modeling in a given application context such as a news recommendation service, it is essential to understand the semantics of Twitter messages. Rowe et al. [3] propose to exploit contextual information in order to clarify the semantics of tweets. In some instances, background information is required in order to utilize user data more effectively. Linked Data principles allow for publishing background information in such a way that the data can be readily consumed by applications[2]. Today, the Linked Open Data (LOD) cloud already provides a great variety of information that can support various applications [4], including expert

[1] http://twitter.com
[2] http://www.w3.org/DesignIssues/LinkedData.html

M. Brambilla, T. Tokuda, and R. Tolksdorf (Eds.): ICWE 2012, LNCS 7387, pp. 378–385, 2012.

finding [5], semantic enrichment of tweets [3], and a rule-based framework for user modeling [6]. Yet, there are, to the best of our knowledge, no research studies that investigate to what extent LOD is beneficial for user modeling processes that analyze user behavior observed on the Social Web.

It should be stressed, that connecting user data with information from the LOD cloud is a challenging task. While the semantics of linked data are well described and facts can easily be retrieved by means of RDF statements, user data on the Social Web often lacks well-defined semantics. Consider Twitter messages as an example: it is easy to extract meta-data such as the creator or creation time of a tweet, but it is challenging to automatically infer the semantic meaning of a tweet. Recently, researchers have begun to make use of named entity recognition services such as OpenCalais[3] and DBpedia Spotlight[4] to infer the topics of Twitter messages, e.g. [2,7].

Understanding the semantics of user data leads to interesting applications such as the profiling of places [7]. Apart from inferring the main location of Twitter users [8], semantic enrichment is also helpful for user modeling and particularly for deducing user interests from Social Web streams [2]. In this paper, we go beyond the aforementioned works and investigate whether background knowledge from the LOD cloud further improves user modeling effectiveness. We analyze our user modeling framework in the context of geographic recommender systems which recommend points of interest (POIs) to users. We explore how Twitter and Flickr can be utilized as user data sources and investigate how background information from GeoNames[5] and DBpedia[6] can be exploited to improve user modeling and consequently the performance of the recommender systems.

The main contributions of our work are as follows: (i) a user modeling framework that exploits the Linked Open Data cloud, (ii) a showcase in which we apply the framework to recommending POIs, and (iii) the evaluation of our methods based on a large Flickr and Twitter dataset which shows the benefits of considering LOD.

2 User Modeling on the Social Web with Linked Data

We now introduce the core building blocks of our user modeling framework. They allow us to exploit Social Web data and knowledge gathered from the LOD cloud to translate user interests into semantic concepts. An overview of our framework is shown in Figure 1. It derives user interest profiles which consist of a set of weighted concepts (each concept is identified by a URI). The concepts are typically dependent on the domain of the application that is requesting user profiles. The weight associated with each concept indicates the intensity of the user's interest in the concept: the higher the weight, the higher the inferred interest. Our

[3] http://www.opencalais.com
[4] http://dbpedia.org/spotlight
[5] http://geonames.org
[6] http://dbpedia.org

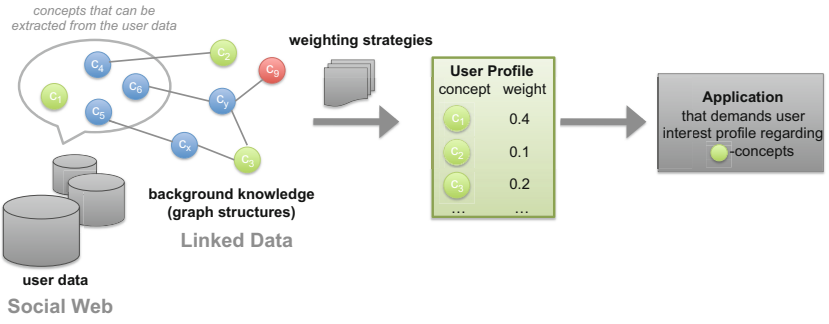


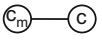
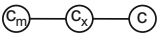
Fig. 1. Overview of the user modeling framework and its three main dimensions: (1) user data sources (what user data to exploit), (2) background knowledge (how to exploit the background knowledge) and (3) weighting strategies (how to weigh the concepts of interests)

framework allows for the creation of various user modeling strategies. Here, we first analyze three design dimensions in detail, namely (i) user data, (ii) background knowledge, and (iii) weighting strategies (see Figure 1). Then, we employ our framework in the domain of geospatial-centric user modeling.

User Data. Our framework provides methods for collecting a user's data from different Social Web streams including her Twitter stream and her resource sharing activities on platforms such as Flickr. As part of the framework's semantic enrichment process, meta-data and semantics are extracted from the observed user activities, the latter being achieved via the named entity recognition service DBpedia Spotlight. Extracted concepts are mapped to the corresponding RDF resources (URIs) in the LOD cloud. The framework can thus represent user activities via the meta-data and via the RDF resources that are related to a user activity. For example, a Twitter message such as "Enjoying the view from the Eiffel Tower" can be represented via information about the application from which the user posted the message and via the semantic concept that can be extracted from the message, namely *http://dbpedia.org/resource/Eiffel_Tower*.

Background Knowledge. Given the concepts extracted from the user data, our framework then acquires background information about the concepts. By following the corresponding URIs, RDF statements are collected to gain a better understanding of a user's interests in concepts that matter to the application that requests the user profile. Lets consider the concept graph example depicted in Figure 1. An application may only be interested in a user's preferences wrt. concepts c_1, c_2 and c_3. Based on the semantic enrichment of the user data, our framework can detect that the user was directly concerned with c_1. By exploiting background information and in particular by following the URIs in the LOD cloud, our framework can infer that the user was also concerned with

Table 1. Examples of (RDF) graph patterns that can be applied to relate a concept c_m, which can directly be extracted from the user data, to a concept of interest c

pattern	description
1. 	*direct mention:* a concept of interest c is directly mentioned in the user data
2. 	*indirect mention I:* a concept c_m is mentioned that occurs in an RDF statement with the concept of interest c; possible RDF graph patterns: (a) $\{c_m\ p\ c\}$ and (b) $\{c\ p\ c_m\}$
3. 	*indirect mention II:* a concept c_m is mentioned that is related to the concept of interest c via another concept c_x; possible RDF graph patterns: (a) $\{c_m\ p_1\ c_x.\ c\ p_2\ c_x\}$, (b) $\{c_m\ p_1\ c_x.\ c_x\ p_2\ c\}$, (c) $\{c_x\ p_1\ c_m.\ c\ p_2\ c_x\}$, (d) $\{c_x\ p_1\ c_m.\ c_x\ p_2\ c\}$

concepts that are related to the concepts that matter for the given application. For instance, the user may have mentioned c_4 which is directly related to c_2 or she may have mentioned c_5 which is indirectly related to c_3.

Different graph patterns (which can be formulated by means of SPARQL queries) can therefore yield different policies for the kind of background knowledge that should be considered in the user profile construction process. Table 1 lists the different graph patterns we analyze in this work. They range from (1) direct mentions of concepts of interests to (3) patterns that relate a mentioned concept c_m with a concept of interest c via another concept c_x. For example, (3.a) describes a situation where c_m and c share the same property value c_x.

Weighting Strategies. Our proposed user modeling framework features different strategies for weighting the concepts of interests for which relations can be discovered according to the aforementioned graph patterns. The basic strategy counts the number of occurrences of a concept c_m, which is related to c via some graph pattern, in the user's data stream to determine the weight associated with c. The weights in a user profile are then normalized so that the sum of the weights is equal to 1.

Geospatial-Centric User Modeling. In this work, the actual application that we evaluate our framework on can be described as follows: *Given a set of POIs and a user* u*, a user modeling strategy has to assign to each POI* p *a weight that reflects to what extent* u *is interested in* p.

We rely Twitter and Flickr as *user data* sources and consider only location-related concepts. From the Twitter stream, we extract the semantic concepts (DBpedia URIs) that are related to places (*http://dbpedia.org/ontology/Place*). In the case of Flickr, we employ an approach that estimates the geographic location of images [11]. The extracted geographic concepts are then utilized to create the geo-related interest profile. Considering these two main Social Web platforms, we

have three options of user data sources when creating a user modeling strategy: (1) Twitter, (2) Flickr, (3) Twitter *and* Flickr.

We obtain background information (RDF statements) about the geospatial concepts that are extracted from the user data and about the points of interests for which the application demands user preferences from DBpedia. For relating the concepts from the user data with the POIs, we utilize particularly the following three graph patterns (see Table 1): (1) direct mentions, (2.a) indirect mentions I, and (3.a) indirect mentions II where a mentioned concept c_m and a POI share the same property value.

To assign a preference score to a POI, we apply the occurrence-based weighting strategy. Thus, we count the number of user activities (represented via the extracted semantic concepts) which match a graph pattern that is employed by the user modeling strategy.

Overall, we thus have $3 \times 3 = 9$ different geospatial-centric user modeling strategies *um*: (i) *um(Flickr, direct mentions)*, (ii) *um(Flickr, indirect mentions I)*, etc. Moreover, we experiment with combining different strategies such as *mix(um(Flickr, direct mentions)* and *um(Flickr, indirect mentions I))* where the preference score is defined as harmonic mean of the scores computed by the individual user modeling strategies.

3 Evaluation of Geospatial-Centric User Modeling

In this section, we evaluate the effectiveness of user modeling strategies that are featured in our user modeling framework. We measure the quality of the different user modeling strategies in inferring user preferences for POIs and investigate the following research questions:

1. How does the source of user data influence the quality of predicting user preferences?
2. How does the inclusion of background knowledge from the LOD cloud impact the user modeling quality?
3. Which (combinations of) user modeling strategies yield the highest effectiveness?

3.1 Experimental Setup: Recommending Points of Interests

To answer the research questions above, we test our user modeling strategies in the context of a recommender system that recommends POIs to a user. Given a user u and a candidate set of POIs such as museums or other tourist attractions, the recommender provides a ranking of POIs so that those POIs which are most relevant to u appear at the top of the ranking. The actual recommender algorithm thus orders the POIs according to the preference scores in u's profile which is derived by a user modeling strategy. The recommendation quality thus solely depends on the quality of the user modeling process.

To investigate to what extent user information from more than one Social Web portal can support the recommendation of POIs, we identified 394 users

who have an account on Flickr *and* Twitter. We accumulated eleven months worth of user activities on both streams. On Flickr, these users uploaded a total of 833,441 images, 16.8% of which are geo-tagged. Based on the tags and title terms we were able to derive a location estimate for 473,129 of the remaining 693,456 images that had not been geo-tagged. Details of the approach can be found in [11]. To translate a given (or estimated) latitude/longitude into a DBpedia POI, we relied on the *findNearbyWikipedia* web service[7]. With this approach we were able to identify one or more DBpedia entries within a radius of 10km for a total of 588,092 images (70.6%). On Twitter, the 394 users posted a total of 2,489,088 tweets. For approximately 11% of the tweets we were able to extract geospatial DBpedia concepts.

We rely on precision, recall, and F-measure (within the top k) to quantify the recommender quality. For user modeling and evaluation purposes we split our dataset as follows: we derived user models based on the first 9 months of user activity and evaluated the models on the final two months of the logged user activities. A POI is considered to be relevant for a user u if the POI is spatially closest to a location where the user took a Flickr photo or if the POI was directly mentioned in a tweet that the user posted within these two months. The split resulted in 9916 candidate POIs of which, on average, 59.35 were considered to be relevant for a given user.

3.2 Results

User Data Sources. When comparing the impact different user-data sources (i.e. utilizing Twitter or Flickr or a combination of both) have on the user modeling quality and subsequently the recommendation quality, our results[8] show that Twitter alone is a more valuable source for creating user profiles that feature preferences in POIs than Flickr alone. However, using both Twitter and Flickr as sources for creating user profiles yields the highest effectiveness, indicating that the two user data sources complement each other to some extent, i.e. Twitter-based profiles provide user preferences which cannot be inferred from Flickr activities and vice versa.

Background Knowledge. Table 2 illustrates the effect of each strategy for exploiting background knowledge in order to relate the concepts, which are extracted from the Twitter *and* Flickr activities, to the POIs. While there is no significant difference in performance between the strategy that considers merely *direct mentions* and the strategy that considers merely *indirect mentions I*, we observe that *indirect mentions II*, which relates mentioned concepts and POIs via shared property values, clearly yields the best performance in terms of the precision, recall, and F-measure within the top 10 and top 20 results.

The results presented in Table 2 also reveal that the combination of different graph patterns for inferring the user preferences in POIs further enhances the

[7] `http://www.geonames.org/export/wikipedia-webservice.html`
[8] Due to space constraints, detailed results are omitted.

Table 2. Overview of the different strategies for integrating background knowledge. Twitter *and* Flickr are used in combination as user data source.

strategy	P@10	R@10	F@10	P@20	R@20	F@20
core strategies:						
direct mentions	0.715	0.260	0.412	0.580	0.179	0.298
indirect mentions I	0.699	0.268	0.426	0.566	0.185	0.308
indirect mentions II	**0.820**	**0.312**	**0.475**	**0.727**	**0.436**	**0.569**
combined strategies:						
direct & indirect mentions I	0.733	0.216	0.360	0.608	0.287	0.416
direct & indirect mentions II	0.836	0.333	0.4975	0.747	0.466	0.596
indirect mentions I + II	0.830	0.325	0.489	0.739	0.456	0.587
direct & indirect mentions I + II	**0.839**	**0.337**	**0.502**	**0.751**	**0.473**	**0.603**

quality of the user modeling and recommendation process. When considering the combination of direct mentions and background knowledge derived from graph patterns of the LOD cloud (indirect mentions I + II), we achieve the highest effectiveness across all evaluation measures: $P@10 = 0.84$, $R@10 = 0.34$, and $F@10 = 0.50$ respectively (last row in Table 2). In comparison with the *direct mention* strategy, which does not exploit RDF statements from the LOD cloud, the $F@20$ performance has more than doubled. Thus, we conclude that taking background knowledge obtained from the LOD cloud into account can significantly improve the effectiveness of user modeling on the Social Web.

Furthermore, we can answer the research questions raised at the beginning of this section as follows. For the task of recommending POIs, it turns out that (1) the aggregation of Twitter and Flickr user data yields the best user modeling performance and that (2) the user modeling quality increases when more background information from the LOD cloud is included. Finally, (3) the best performance is achieved by combining the different graph patterns for acquiring background information and inferring user preferences.

4 Conclusions

In this paper, we proposed a framework for enriching user modeling on the Social Web with information from the Linked Open Data cloud. Our framework monitors user activities on Social Web platforms such as Twitter and Flickr, infers the semantic meaning of user activities and provides strategies for gathering background information from the Web of Data to generate semantically meaningful user profiles that support a given application. We showcased and evaluated our framework in the context of a geospatial recommender system where the core challenge lies in deducing user preferences for POIs. To account for this, we also presented a method that allows for the semantic enrichment of Flickr pictures by (i) estimating the geographical location where a picture was taken and by (ii) exploiting GeoNames in order to identify related DBpedia concepts.

Our evaluation showed the effectiveness of our user modeling framework. Based on a large Twitter and Flickr dataset of more than 2.4 million tweets and 800 thousand Flickr pictures that we obtained by monitoring 394 users over a period of nearly a year, we revealed that the aggregation of user data from both Social Web platforms is beneficial for inferring user preferences. Taking advantage of background information derived from the LOD cloud led to substantial improvements of the baseline user modeling effectiveness.

Acknowledgements. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no ICT 257831 (ImREAL project).

References

1. Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and tweet: experiments on recommending content from information streams. In: Proc. of the 28th Int. Conf. on Human Factors in Computing Systems(CHI), pp. 1185–1194. ACM (2010)
2. Abel, F., Gao, Q., Houben, G.J., Tao, K.: Analyzing User Modeling on Twitter for Personalized News Recommendations. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) UMAP 2011. LNCS, vol. 6787, pp. 1–12. Springer, Heidelberg (2011)
3. Rowe, M., Stankovic, M.: Aligning Tweets with Events: Automation via Semantics. The Semantic Web Journal, Special Issue on Interoperability, Usability, Applicability (2011)
4. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Int. Journal on Semantic Web and Information Systems (IJSWIS) 5(3), 1–22 (2009)
5. Stankovic, M., Wagner, C., Jovanovic, J., Laublet, P.: Looking for Experts? What can Linked Data do for You? In: Workshop on Linked Data on the Web (LDOW), Raleigh, USA (2010)
6. Leonardi, E., Abel, F., Heckmann, D., Herder, E., Hidders, J., Houben, G.-J.: A Flexible Rule-Based Method for Interlinking, Integrating, and Enriching User Data. In: Benatallah, B., Casati, F., Kappel, G., Rossi, G. (eds.) ICWE 2010. LNCS, vol. 6189, pp. 322–336. Springer, Heidelberg (2010)
7. Cano, A.E., Varga, A., Ciravegna, F.: Volatile Classification of Point of Interests based on Social Activity Streams. In: Workshop on Social Data on the Web (SDoW), Bonn, Germany (2011)
8. Hecht, B., Hong, L., Suh, B., Chi, E.H.: Tweets from Justin Bieber's Heart: The Dynamics of the "Location" Field in User Profiles. In: Proc. of Int. Conf. on Human Factors in Computing Systems (CHI), Vancouver, BC, Canada. ACM (2011)
9. Golbeck, J., Hansen, D.L.: Computing Political Preference among Twitter Followers. In: Proc. of Int. Conf. on Human Factors in Computing Systems (CHI), Vancouver, BC, Canada. ACM (2011)
10. Pennacchiotti, M., Popescu, A.M.: A Machine Learning Approach to Twitter User Classification. In: Proc. of the 5th Int. AAAI Conf. on Weblogs and Social Media (ICWSM), Barcelona, Spain. AAAI Press (2011)
11. Hauff, C., Houben, G.-J.: Geo-Location Estimation of Flickr Images: Social Web Based Enrichment. In: Baeza-Yates, R., de Vries, A.P., Zaragoza, H., Cambazoglu, B.B., Murdock, V., Lempel, R., Silvestri, F. (eds.) ECIR 2012. LNCS, vol. 7224, pp. 85–96. Springer, Heidelberg (2012)