Extraction of Professional Interests from Social Web Profiles

Fabian Abel[1], Eelco Herder[2], Daniel Krause[2]

[1] Web Information Systems, TU Delft, The Netherlands
f.abel@tudelft.nl
[2] IVS – Semantic Web Group & L3S Research Center, Leibniz University Hannover, Germany
{herder,krause}@l3s.de

Abstract. Many people share and communicate their private thoughts and opinions via systems like Facebook and Twitter. In this paper, we analyze if also professional interests of a user can be extracted from these activities and be distinguished from private interests. The results indicate that performance largely depends on the size and quality of the Social Web profiles. Methods for reducing noise and chatter for-high volume profiles improve quality, but reduce diversity of the profiles.

1 Introduction

Today, one can observe a paradigm shift on the Web from a rather machine-centered view towards a more user and community-centered view. The term *Social Web* describes this paradigm shift and highlights a new culture of participation on the Web. As part of their daily routines, people share and communicate their thoughts and opinions via Social Web systems like Facebook[3] or Twitter[4]. Therefore, a huge amount of social data is available on the Web. For example, people publish more than 65 million messages on Twitter per day[5] about diverse topics, ranging from casual chatter to news. This large and multifaceted reservoir of social data promises benefits for various applications – in particular for systems that rely on information about their users.

Modeling interests and concerns of users has been studied already for different Social Web systems like Delicious [1], Last.fm [2] or Twitter [3]. Furthermore, there exist studies on cross-system user modeling [4] and cross-system user modeling on the Social Web in particular [5]. However, correlations between the user's professional and private activities in the real world and her behavior in Social Web systems have not been researched extensively yet. For example, are there any correlations between a user's professional interests and her posting behavior on Twitter? Is it possible to distill a user's professional interests from Twitter, Delicious or LinkedIn and which Social Web system is the best source to infer these interests?

[3] http://facebook.com
[4] http://twitter.com
[5] http://techcrunch.com/2010/06/08/twitter-190-million-users/

In this paper, we answer these questions and analyze how professional interests of a user relate to her Social Web activities on Twitter, Delicious and LinkedIn. We investigate strategies for enriching the semantics of social data and study strategies for filtering social data based on interactions with other users in these systems.

2 Related Work

Any adaptation and personalization of Web systems requires user profiling, which involves the collection and interpretation of data and information about a user [6]. The Web 2.0 offers simple interfaces for Internet users to contribute user generated content – an ideal setting for performing user modeling. Especially tagging systems, such as Delicious, Flickr and Last.fm, offer the opportunity to derive tag-based user profiles. Inspired by swarm intelligence of ant colony optimization, Michlmayr et al. [1] present their Add-a-Tag approach which builds a user profile from given Delicious tags and is able to detect change of user's interests over time and to adapt the profile accordingly. Firan et al. [2] used a tag-based user profile generated from tag-assignments of Last.fm. They were able to use the tag profile as input for search algorithms in order to generate music recommendations which outperformed classical collaborative filtering methods. In microblogging applications, like Twitter, users do not contribute keywords but short text snippets. Due to heavy use of abbreviations, noise and various languages, the semantics behind a Tweet need be captured before it can be used as input for user modeling. Therefore, Castillo et al. [7] try to classify tweets as *news* or *chatter* in order to estimate the credibility of information on Twitter. Abel et al. [3] extract topics and entities from Tweets for recommending news articles. They show that the additional processing outperforms hashtag-based profiles.

Berkovsky et al. [8] have shown that the accuracy of a user profile of a recommender system can be increased by integrating data from other recommender systems. As users often have accounts on different Web 2.0 platforms, the combination of user profiles from different platforms might increase the quality of a user profile as well. Abel et al. [5] combined form-based and tag-based Web 2.0 profiles. The tag-based profiles were extracted from Flickr, Delicious, and StumbleUpon and could not only successfully overcome the cold-start problem but also improve the quality of comprehensive single-platform-based profiles.

A remarkable observation is that the authors detected a very small overlap of the tags that a users used in different systems during the combination of the profiles. This leads to the assumption that users use different systems for different purposes: while users use Last.fm mostly to listen to music, users on Facebook connect to their friends; LinkedIn users connect to their business partners, and Twitter users might tweet about both private and business related topics. While the related work outlines that Web2.0 user profiles serve well to predict leisure interests, we are not aware of any work conducted to distinguish leisure and professional activities and predict professional interests of a user.

3 User Modeling for Mining Professional Interests

Our main goal is to extract the professional scientific interests of people from their Social Web interactions. In our first experiment, we use the scientific publications of a group of researchers as a ground truth and try to discover scientific interests of these researchers from the different Social Web platforms that they use. In the second experiment, we use these different Social Web profiles to recommend relevant publications.

3.1 Dataset

For our experiments, we used a dump of the Social Handle Archive[6] (SOHARC). The dump consists of records of 99 persons, containing demographic data, contact information, and usernames of fourteen Web 2.0 platforms, like Twitter, Delicious, and LinkedIn. As the user had the choice which account information to provide, not all profiles were filled completely. In the dataset, we have 78 users who specified a Twitter account, 48 having a LinkedIn profile, 46 user with a Delicious account, and 22 users who filled all three profiles. We used the subset of 32 SOHARC users with at least two of these profiles and for whom we could retrieve the publication data.

We used Mypes[7] to extract the public profile information from LinkedIn and Delicious. For LinkedIn, which provides form-based data, we used the bag-of-word approach to create user profile vectors; for Delicious, we directly used the tags to create the profile vectors. For Twitter, we crawled overall 28.293 Tweets that a) have been created by a SOHARC user, b) retweet a SOHARC user or c) mention a SOHARC user; we applied the bag-of-word approach to generate the user profiles. In a second step, we passed the Twitter profiles to OpenCalais[8] and extracted entities from the Tweets for constructing entity-based Twitter profiles. Finally, we aggregated the Delicious and LinkedIn profiles with a) the bag-of-word-based Twitter profiles and b) the entity-based Twitter profiles.

To extract the professional interests of the SOHARC users, we connected them to their publication records, which are assumed to reflect their professional interests and activities. The publication records are extracted from the STELLAR Open Archive[9]. The publication data includes the title, abstract and keywords of publications. We related the publications to the SOHARC users by manually created mappings, which resulted in 730 assigned publications.

3.2 Overlap of Social Web Profiles with Scientific Profiles

Our main goal is to extract the professional scientific interests of people from their Social Web interactions. To gain first insights into this task, we first analyze how the terminology people use on the Social Web overlaps with the terminology of their scientific publications. In Figure 1 we therefore plot for each user u the fraction of u's publications that feature at least one term (excluding stopwords)

[6] http://soharc.upb.de/

[7] http://mypes.groupme.org

[8] http://www.opencalais.com/

[9] http://oa.stellarnet.eu/ - there is a known overlap between the community of SOHARC users and the authors represented in the STELLAR Open Archive.

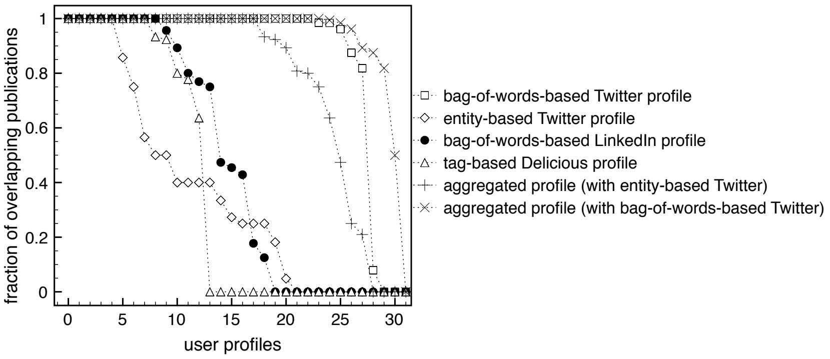

Fig. 1. Fraction of users' publications for which terms from the title, abstract or keywords overlap with the corresponding Social Web user profile. 32 users have co-authored at least one scientific publication. The x-axis shows the x^{th} user starting from 0 to 31.

in the title, abstract or keywords that also occurs in the different Social Web profiles of u.

The Twitter-based profile, which models a user by means of a bag of words, features much higher overlap than the LinkedIn or Delicious profiles (Figure 1). However, the entity-based Twitter user modeling - which extracts the important concepts (entities) from the bag-of-words-based Twitter - reveals that, in fact, entities overlap just a little with the entities mentioned in the title, abstract or keywords of a user's publications. The bag-of-words-based Twitter strategy succeeds in relating 29 of 32 users to their publications, while the entity-based approach succeeds for 21 users. Furthermore, the fraction of publications per user that overlap with the user's profile is much higher for the bag-of-words-based strategy (87%) than for the entity-based Twitter strategy (36%), Delicious strategy (38%) or LinkedIn strategy (46%).

The user modeling strategies that aggregate a user's profile from Twitter, Delicious and LinkedIn increase the performance regarding the overlap with the user's publications. For example, the profile aggregation strategy that combines the entity-based Twitter profiles with the corresponding Delicious and LinkedIn profiles relates, on average, 94% of the publications to the users, when examining terms that appear in both the aggregated profile and the title, abstract or keywords of a publication.

3.3 Recommending Publications based on Social Web Profiles

The positive findings on overlap between the Social Web profiles and the scientific interests of a user motivate our second experiment: now, we aim to recommend relevant publications to SOHARC users based on their Social Web profiles.

We represent each publication as a bag-of-words vector, using the title, keywords and abstract. The user profiles are generated as in our first experiment. The recommendations are based on cosine similarity between the user profile vectors and the publication vectors. As ground truth we use the co-authorship relationship.

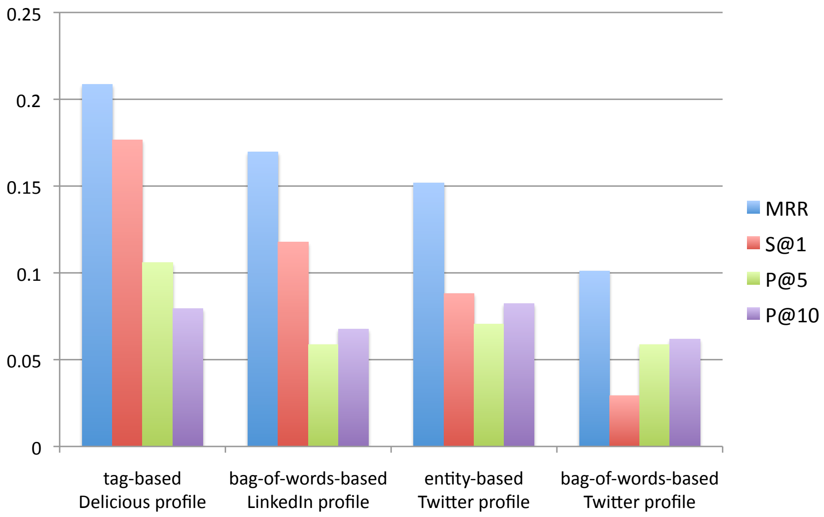

Fig. 2. Performance of user modeling strategies for recommending publications.

Figure 2 shows the performance of the recommendations, as measured by the mean reciprocal rank (MRR), success at rank 1 (S@1) and precision at rank 5 and 10 (P@5 and P@10). According to the MRR, which is the inverse rank of the first hit, and S@1, Delicious outperforms LinkedIn, which on its turn outperforms Twitter. However, regarding P@10, the entity-based Twitter strategy outperforms both Delicious and LinkedIn. Our hypothesis is that Twitter profiles seem to cover more (or a broader range of) professional interests, but also seem to be much more noisy. This is supported by the fact that the semantically enriched entity-based Twitter user profiles perform better than the bag-of-word profiles. By contrast, Delicious profiles (which allow for good quality at high ranks – cf. S@1, MRR) are more focused than Twitter profiles, but they do not cover the whole variety of professional interests. This can also be observed from the drop of P@k for Delicious profiles, which is stronger than the drop for Twitter profiles.

4 Discussion and Conclusion

In this paper we discussed two experiments in which we investigated the presence of professional interests in social Web profiles and how these interests can be used for recommending publications. The results indicate that Delicious and LinkedIn profiles contain only a small amount of selected and specific interests, which cover only a small part of the professional interests. In the second experiment we have seen that these selected profiles provide qualitatively good recommendations, but fail to cover diversity. On the other hand, Twitter profiles generally contain more facets, but suffer from noise.

Obviously, there is a strong correlation between the size of the Social Web profiles and the quality of the extracted professional interest profiles ($p < 0.01$ for all individual social Web profiles and the aggregated profiles). In other words, extraction of professional interest profiles works best for active users of social media (which also indicates that more active use typically coincides with less chatter).

We found that one way to remove chatter or noise from social Web profiles is to extract entities from the keywords. An alternative approach would be to consider only Tweets that have been retweeted by other users - assuming that users typically retweet meaningful or important messages rather than chatter. This assumption is supported by the relatively strong cosine similarity between tweet and retweet profiles (0.36) and the low similarity of retweet profiles with reply profiles (replies have been found often to be chatter, including thank-you messages). However, a preliminary evaluation showed that the potential higher quality of retweet profiles does not compensate for their significantly smaller average size (15% of the size of an average full Twitter profile).

In summary, our results confirm that professional interests can be extracted from social Web profiles. The performance of the extraction largely depends on the sizes of these profiles. Techniques such as entity extraction help to reduce the amount of noise or chatter, but this comes at the price of smaller, less diverse profiles.

References

1. Michlmayr, E., Cayzer, S., Shabajee, P.: Add-A-Tag: Learning Adaptive User Profiles from Bookmark Collections. In: Proc. of the 1st Int. Conf. on Weblogs and Social Media (ICWSM '06). (March 2007)
2. Firan, C.S., Nejdl, W., Paiu, R.: The Benefit of Using Tag-based Profiles. In: Proc. of 2007 Latin American Web Conference (LA-WEB '07), Washington, DC, USA, IEEE Computer Society (2007) 32–41
3. Abel, F., Gao, Q., Houben, G.J., Tao, K.: Analyzing User Modeling on Twitter for Personalized News Recommendations. In: International Conference on User Modeling, Adaptation and Personalization (UMAP), Girona, Spain, Springer (July 2011)
4. Mehta, B.: Learning from what others know: Privacy preserving cross system personalization. In Conati, C., McCoy, K.F., Paliouras, G., eds.: User Modeling. Volume 4511 of Lecture Notes in Computer Science., Springer (2007) 57–66
5. Abel, F., Herder, E., Houben, G.J., Henze, N., Krause, D.: Cross-system user modeling and personalization on the social web. User Modeling and User-Adapted Interaction (UMUAI), Special Issue on Personalization in Social Web Systems (2011) 1–42
6. Brusilovsky, P., Millán, E.: User models for adaptive hypermedia and adaptive educational systems. In Brusilovsky, P., Kobsa, A., Nejdl, W., eds.: The adaptive web. Springer-Verlag, Berlin, Heidelberg (2007) 3–53
7. Castillo, C., Mendoza, M., Poblete, B.: Information Credibility on Twitter. In: WWW '11: Proceedings of the 20th international conference on World wide web, Hyderabad, India, ACM Press (2011)
8. Berkovsky, S., Kuflik, T., Ricci, F.: Mediation of user models for enhanced personalization in recommender systems. User Modeling and User-Adapted Interaction (UMUAI) **18**(3) (2008) 245–286