TwikiMe!

User profiles that make sense.

Patrick Siehndel, Ricardo Kawase

L3S Research Center, Leibniz University Hannover, Germany
{siehndel, kawase}@L3S.de

Abstract. The use of social media has been rapidly increasing in the last years. Social media, such as Twitter, has become an important source of information for a variety of people. The public availability of data describing some of these social networks has led to a great deal of research in this area. Link prediction, user classification and community detection are some of the main research areas related to social networks. In this paper, we present a user modeling framework that uses Wikipedia to model user interests inside a social network. Our model of user interests reflects the areas a user is interested in, as well as the level of expertise a user has in a certain field.

1 Introduction

In the last years, surfing on social network Web sites has became the most prominent online activity. Together, the top accessed social networks such as Facebook, Twitter and Myspace, to name a few, aggregate over a billion users. Given this level of activity, the research interest on the field of social networks has grown considerably. User modeling, link prediction, sentiment analysis, community analysis, sociology and many other areas of Web Science are examples of research fields exploiting the public (and private) data available in such networks.

In this paper we propose a semantic approach to generate user profiles based on the publicly available Twitter data. In other words, we generate a concise, yet descriptive semantic user profile using Twitter streams. With a semantically generated user profile, one can easily identify the exact topics of interest a user has. In contrast to bag of word approaches, we generate semantically enhanced user profiles that quantify the users' interests in a set of specific categories. Untimately, our profiling method outputs a semantically enhanced user profile that reflects the real user interest (based on his explicit tweets). The TwikiMe! user profile is a reduced representation of the user interests on a 23-sized vector that is both human and machine comprehensible.

2 Twikifying

We provide TwikiMe! as a framework to generate semantically enhanced profiles for Twitter users. To accomplish it, we use the well established Wikipedia corpus as a semantic knowledge base. Wikipedia is arguably the most accessed reference Web site and each of the over 3.5 million existing articles are manually classified by human

curators to one or more categories. Additionally, categories are organized in a graph where sub-categories reference top level categories. The English Wikipedia has in total 23 top level categories (*Main topic classifications*), which we use to represent a user profile (See Figure 2 for the list of top-level categories). Abel et al. [2] presented similar strategies to enhance Twitter user profiles, however their topic-based profile is built upon topics related to different types of news events. In our work, we consider the topics (categories) of each detected Wikipedia entity, thus the categories describe a wider area of fields.

The creation of semantically enhanced user models consists of three stages (see Figure 1). The first stage, *Extraction*, entities are extracted from the user's tweets. Given Twitter users streams, we annotate all tweets to detect any mention of entities that can be linked to Wikipedia articles. In this step, we chose to use the WikipediaMiner [1] service to annotate the users' tweets.

The second stage, *Categorization*, we extract the categories of each entity that has been mentioned in the users Tweets. For each category, we follow the path of all parent categories, up to the root category. In some cases, this procedure results in the assignment of several top level categories to an entity. A weight is calculated for each category by first setting a value of 1 for the detected entity. Following the parent categories (which are closer the root category) we divide the weight of each node by the number of siblings categories, resulting in each entity receiving 23 categories' scores. In the final stage, *Aggregation*, we perform a linear aggregation over all of the scores for a tweet in order to generate the final user profile.

Our approach of reducing the user profile to 23 topics differs significantly from the work of Michelson and Macskassy [4]. In their work, they propose a similar approach to annotate tweets with Wikipedia articles; but instead, of considering all parent categories, they traverse the category graph only "5 levels deep". In doing so, they assume that a five stage traversal is sufficient to reach categories that are general enough for a user's profile. The limitation of their assumption is that a user's classification may have an unlimited number of categories, thereby preventing profiles from having a normalized length and comparison among all users.

We applied our user profiling method on the Twitter dataset created by Yang and Leskovec [5] containing 476 million tweets from 20 million users. The tweets where collected between 1st of June 2009 and 31st of December 2009. We combined with the social Graph created by Kwak et al. [3] with 41.7 million user profiles and 1.47 billion social relations. We randomly selected 20,000 users with their followers and friends for a total of 630,000 users and 28 million tweets and analyzed how similar the users are to their followers and friends

Additionally, we calculated how much the user profile changes over time by splitting the user's tweets into two bins and calculating the inter-user similarity. To perform such evaluation we divided the users' stream in half (by number of tweets), comparing the profile of the earlier tweets against the latest ones. Table 1 shows the results of the similarity measures. As metric we used the cosine similarity.

Our result show that while the categories related to the tweets of one user stay the same over the time (leading to a very high similarity of 0.85), we see that the average similarity between the user and her friends and followers is rather low with 0.26 and

Table 1. Cosine Similarity of the generated user profiles

Comparison	User - User	User - Followers	User - Friends	Friends - Followers
Cosine-Similarity	0.85	0.27	0.26	0.24

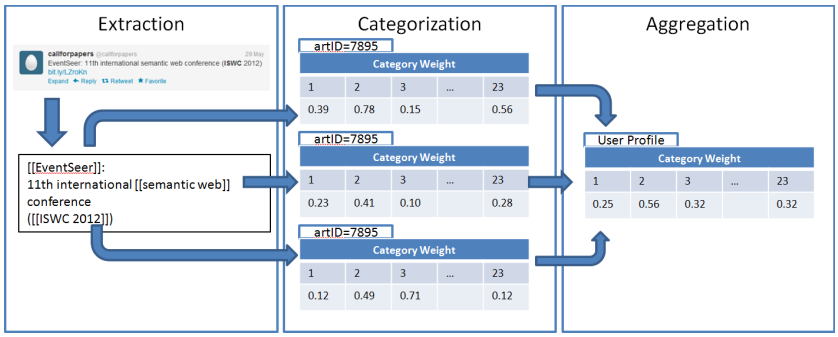

Fig. 1. Different stages for the creation of userprofiles

0.27. Comparing the profiles of the followers and friends we obtained a similarity of 0.24. Based on the high similarity when comparing only tweets from one user, we assume that the topics one user tweets about stays rather constant, and are not necessarily the same as the topics discussed by friends and followers.

3 TwikiMe! Online

To ilustrate our approach, we deployed an online version[1] of the TwikiMe! framework. The interface guides the user through the twikifying process. First, one must provide a valid Twitter user name and the amount of past tweets to be annotated (due to the outsourcing of WikipediaMiner service, this step is the most time consuming). Once the process is finished, the user is asked to trigger the Categorization step. Finally, the user is presented with the TwikiMe! chart of the given Twitter user.

To address qualitative comparisons, we also provide a comparative graphical interface as depicted in Figure 2. In the comparative chart, it is possible to visualize the percentage results for two distinct users, side-by-side. Additionally, the interface provides a date picker to restrict the comparisons to a given time interval. In Figure 2, one can see the comparison between the Twitter accounts of the Dalai Lama (@DalaiLama) and president Barack Obama (@BarackObama), during the period of April through December 2011. The resulting user profile depicted in the chart shows a clear dominance of Barack Obama in the topics of "people", "politics" and "business", while Dalai Lama leads on "life", "culture" and "geography" topics.

A closer look into different "known" users (politicians, researchers and technology news), shows that the resulting TwikiMe! profiles provide expected topic identities. While technology news providers lead on the topics of "business", "society" and "technology", researchers lead on the topics of "education" and "applied sciences". We invite

[1] `http://twikime.l3s.uni-hannover.de`

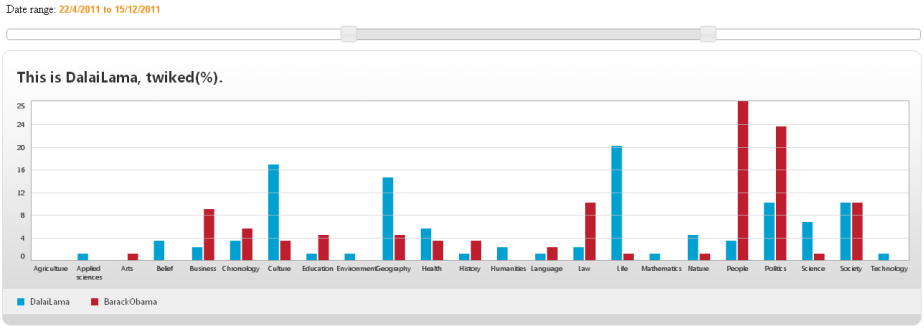

Fig. 2. TwikiMe comparison charts showing the profile of Dalai Lama and Barack Obama.

the readers to try it by themselves our TwikiMe! prototype at `http://twikime.l3s.uni-hannover.de`. The prototype makes available all previously *twikified* users for comparisons.

4 Conclusion

In this paper, we introduced the TwikiMe! A prototype for generating comprehensible user profiles that are represented compactly as a 23-length vector. The profiles quantifies the users' implicit interest in each of these different categories. We believe that, by exploiting the TwikiMe! profiles, we are able to improve content and user recommendation on Twitter.

As future work, we plan to empirically demonstrate the proof of concept of the generated user profiles as well as improving the strategy of the Categorization step. Finally, we believe that there may be a significant difference between what a user "produces" and what a user "consumes". We also plan to evaluate this difference by generating TwikiMe! user profiles based on tweets that a user follows, rather than just the tweets that a user writes.

References

1. Wikipediaminer toolkit. website, 2012. http://wikipedia-miner.cms.waikato.ac.nz/ (accessed july 30, 2012)., 2011.
2. F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing User Modeling on Twitter for Personalized News Recommendations. In *International Conference on User Modeling, Adaptation and Personalization (UMAP), Girona, Spain*. Springer, July 2011.
3. H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.
4. 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*, AND '10, pages 73–80, New York, NY, USA, 2010. ACM.
5. J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 177–186, New York, NY, USA, 2011. ACM.