

Singapore Management University Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems

School of Information Systems

12-2014

Extracting Interest Tags from Twitter User Biographies

Ying DING


Singapore Management University, ying.ding.2011@smu.edu.sg

Jing JIANG

Singapore Management University, jingjiang@smu.edu.sg

DOI: https://doi.org/10.1007/978-3-319-12844-3_23

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the [Databases and Information Systems Commons](#), [Numerical Analysis and Scientific Computing Commons](#), and the [Social Media Commons](#)

Citation

DING, Ying and JIANG, Jing. Extracting Interest Tags from Twitter User Biographies. (2014). *Information Retrieval Technology: 10th Asia Information Retrieval Societies Conference, AIRS 2014, Kuching, Malaysia, December 3-5, 2014: Proceedings*. 8870, 268-279.

Research Collection School Of Information Systems.

Available at: https://ink.library.smu.edu.sg/sis_research/2635

This Conference Proceeding Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

Extracting Interest Tags from Twitter User Biographies

Ying Ding and Jing Jiang

School of Information Systems, Singapore Management University, Singapore
{ying.ding.2011, jingjiang}@smu.edu.sg

Abstract. Twitter, one of the most popular social media platforms, has been studied from different angles. One of the important sources of information in Twitter is users' biographies, which are short self-introductions written by users in free form. Biographies often describe users' background and interests. However, to the best of our knowledge, there has not been much work trying to extract information from Twitter biographies. In this work, we study how to extract information revealing users' personal interests from Twitter biographies. A sequential labeling model is trained with automatically constructed labeled data. The popular patterns expressing user interests are extracted and analyzed. We also study the connection between interest tags extracted from user biographies and tweet content, and find that there is a weak linkage between them, suggesting that biographies can potentially serve as a complimentary source of information to tweets.

1 Introduction

With a large percentage of the population, especially youngsters, using social media to communicate with families and friends nowadays, much personal information such as a user's gender, age and personal interests has been revealed online. Such personal information is of great value to both research and industry. For example, social scientists and psychologists can better study people's behaviors based on the tremendous amount of user information collected from social media [11], [16]. Companies can do targeted online advertising more accurately based on users' personal information. While personal information can be mined from various sources, an important source that has been largely neglected is users' biographies written by themselves in social media platforms. In this work we study how we can extract users' personal interests from Twitter user biographies.

In many social media platforms such as Twitter, users are given the opportunity to describe themselves using one or two sentences, which we refer to as biographies. Compared to structured user profiles in Facebook and LinkedIn, biographies are often written in free form, which is hard for computers to understand. However, because these biographies are written by users themselves, they are expected to reflect the users' background, interests and beliefs. For example, Table 1 shows a few example biographies from Twitter. We can see that in the first and the second biographies, there are phrases such as "Soccer fan" and "love video games" that clearly describe a user's interests. Phrases such as "Software Prof" shows a user's profession and "18 year old" shows a user's age. The third biography, on the other hand, is a popular quote "live life to its

fullest,” which reveals the user’s attitude to life. A recent study based on a sample of users found that around 28% of Singapore Twitter users and over 50% of US Twitter users revealed personal interests in their biographies, which suggests the high value of mining Twitter biographies [4].

Table 1. Examples of user biographies on Twitter

User #1	Hyderabadi Ladka, Software Prof, Soccer fan
User #2	18 year old, theatre kid I love video games and cooking. Hmu on SnapChat (anonymaxx) Instagram @MaxxReginello
User #3	Young wild’n free... Hahaha, live life to the fullest with not regrets..

Intuitively, if we can automatically extract phrases such as “soccer” and “video games” from user biographies, these phrases serve as meaningful and informative tags for the respective users. We refer to this kind of words and phrases that describe users’ personal interests as *interest tags*. In this paper, we try to automatically extract these interest tags. We also try to link the extracted interest tags to users’ content (tweets) and study their correlations.

Extracting interest tags can be treated as a typical information extraction problem. A ready solution is to employ supervised sequence labeling algorithms such as CRF with labeled training data. However, manual annotation to obtain labeled data is labor-intensive. We observe that there are a few common syntactic patterns people use to describe their interests. We therefore first build a noisy training data set using a set of seed patterns and heuristic rules, and then train a CRF model on this labeled data set. With this approach, we are able to achieve an F1 score of 0.76 for interest tag extraction. We also show some top interest tags as well as pattern words found in our data set.

While a Twitter user may describe her interests in her biography, these interests may not be clearly reflected in her published tweets. To understand whether and which interest tags are likely to be represented by users’ tweets, we further study the connection between interest tags and Twitter content.

The contributions of our work can be summarized as follows. We are the first to study how to extract interest tags from user biographies in social media. We show that with state-of-the-art information extraction techniques, we can achieve decent performance for this task. We also show that not all interest tags are reflected in users’ tweets, which suggests that it is not sufficient to only consider tweets for finding user interests. We expect that interest tags extracted from user biographies can potentially be used for user profiling and many other applications.

The rest of the paper is organized as follows. We first review related work in Section 2. Then we present our observation about linguistic patterns of interest tags in Section 3. Our method and experiments will be presented in Section 4 and Section 5. We analyze and discuss the connection between interest tags and tweet content in Section 6. We conclude this work and give suggestions on future work in Section 7.

2 Related Work

Our work is related to a few different lines of work, which we briefly review below.

2.1 Psychological Studies on User Profiles

Biographies in social media have attracted much attention in the psychology community. Different questions have been studied in recent years. Counts and Stecher studied the creation process of profiles and found that people create profiles to match their self-representation profiles [3]. Lampe *et al.* study the relationship between profile structure and number of friends and discovered that some of the profile elements can predict friendship links [9]. Disclosure of user information in biographies is also an important problem to study in social media [14]. Profile information has also been used to do prediction in other applications, such as user need prediction [22] and sensational interest prediction [7]. However, studies in psychology do not focus on computational methods to automatically extract information from user biographies. In contrast, our work is about automatic extraction of user interests from biographies.

2.2 User Profile Construction

User profiling is an important research question, which aims at extracting and inferring attributes of a user from all his/her online behaviour. There has been a number of studies on user profile construction from different angles. Roshchina *et al.* built user profiles according to personality characteristics mined from review text [20]. They showed that based on a large number of reviews, it is possible to differentiate personality types and match users with reviews written by people with similar personality. Pennacchiotti and Popescu proposed a general framework to build user profiles [15]. Their work combines classification algorithms and label propagation in social network graphs. Their method shows encouraging results on three different tasks, which are identification of political affiliation, ethnicity and business affinity. Demographics information is one of the most important aspects in a user's profile. Gender and age prediction has also attracted much attention and has been studied in some recent work [2], [5], [12,13], [19].

User interest is also an important type of information for profiling a user. In this work, we extract users' interests from their biographies, which has not been done in existing work on user profile construction.

2.3 Information Extraction in Social Media

With the explosion of content generated in social media, information extraction, which aims at extracting structured, meaningful information from unstructured, noisy content edited by online users, becomes more necessary than ever before. Ritter *et al.* proposed an open domain event extraction approach, which leverages the large volume of Twitter messages and outperforms a supervised baseline [18]. Ritter *et al.* designed a new pipeline of named entity recognition adapted to Twitter corpus [17]. Benson *et al.* utilized a supervised principled model to extract events-related information from social media feeds [1]. A similar study was done by Imran *et al.* [8], which focuses more on valuable information of disasters.

Extracting tags to represent users’ interests is very helpful to various applications such as online advertisement, friend recommendation. Wu *et al.* applied two standard methods (TF-IDF and TextRank) on tweets to generate personalized tags for users [21]. Liu *et al.* solved tag generation problem with a machine translation technique [10]. However, no work has systematically looked at information extraction from user biographies in social media.

3 Linguistic Patterns of Interest Tags

In order to design a suitable method to extract interest tags from user biographies, we need to first understand how users typically describe their interests in biographies. In this section, we show some typical linguistic patterns we found from a sample of Twitter biographies.

One author of this paper took a random sample of 500 biographies from our Twitter data (described in Section 5) and manually examined them carefully. First of all, the examination revealed that only 28.8% of biographies contain meaningful interest tags. The rest of the biographies often describes the user’s attitude, belief or other demographic information. Among the biographies that do contain interest tags, the author found a set of common patterns as shown in Table 2. We can see that the majority of interest tags are expressed by the “Noun + Noun” pattern, where the first noun (or noun phrase) is the interest tag.

Table 2. Syntactic patterns describing user interests in Twitter biographies

Relative Frequency	Pattern	Example
66%	Noun + Noun	football fan
3%	Noun/Verb + Prep + Noun	obsessed with football / fan of football
5%	Verb + Noun	love football
26%	Others	reader / Beliebers should follow me. (Note: “Belieber” here means a fan of Justin Bieber.)

It is worth pointing out that in all these patterns, the interest itself is described by a noun or noun phrase such as “football” and “video games” and the rest of the pattern can be regarded as some kind of *trigger*. For example, words such as “fan” and “junkie” strongly indicate that there is a word or phrase nearby that describes an interest. This observation shows that contextual words are useful for identifying interest tags, which will guide our design of features when we apply supervised learning for tag extraction.

4 Extracting Interest Tags

Based on the analysis as described in the previous section, we design our solution in the following way. Our method eventually uses conditional random fields (CRFs), which represents the state of the art of information extraction. It defines the conditional probability of a sequence label $y_{1:N}$ given the observation sequence $x_{1:N}$ as

$$p(y_{1:N}|x_{1:N}) = \frac{1}{\mathbf{Z}} \exp\left(\sum_i \sum_k \lambda_k f_k(y_{i-1}, y_i, x_{1:N})\right), \quad (1)$$

where \mathbf{Z} is the normalization factor that can be calculated as

$$\mathbf{Z} = \sum_{y_{1:N}} \exp\left(\sum_i \sum_k \lambda_k f_k(y_{i-1}, y_i, x_{1:N})\right). \quad (2)$$

Given a set of training data $\{\mathbf{x}, \mathbf{y}^*\}$, where \mathbf{x} is the observation sequence and \mathbf{y}^* is the correct sequence label, we can learn the model parameter $\Theta = \{\lambda_k\}_1^K$ as follows

$$\Theta = \arg \max_{\Theta} \sum_j \log p(\mathbf{y}_j^* | \mathbf{x}_j, \Theta) + \alpha \sum_k \lambda_k^2, \quad (3)$$

where $\sum_k \lambda_k^2$ is a regularization term. After learning Θ , given an observation sequence \mathbf{x} , we can infer its sequence label $\hat{\mathbf{y}}$ as follows

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} p(\mathbf{y} | \mathbf{x}, \Theta). \quad (4)$$

We model a biography as a sequence of words where each word has a hidden label. Following the common practice of using BIO notation in sequence labelling, each word in a sequence can be assigned one of the following labels: {B-INT, I-INT, O} where B-INT and I-INT indicate the beginning of and inside of an interest tag, respectively, and O indicates outside an interest tag. Fig. 1 shows an example sentence with the labels for each word.

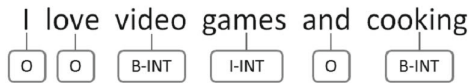


Fig. 1. Example of a biography segment and the corresponding labels

In the rest of this section, we first describe how we obtain labeled data using seed patterns and heuristic rules. We then describe the features we use for CRF.

4.1 Generating Labeled Training Data

We build a noisy labeled data set here to avoid the time-consuming human annotation process. Our goal is to ensure high precision for such automatically generated labeled data. We first perform POS tagging on all sentences of biographies using Twitter-POS-Tagger [6]¹. As we only treat noun phrases as candidate interest tags, we need to identify noun phrases. We heuristically treat a consecutive sequence of nouns as a noun phrase. To get a set of labeled biographies, we start from a set of seed patterns and extract noun phrases inside these patterns. The seed patterns we use are “play + NP,” “NP + fan” and “interested in + NP,” where NP stands for a noun phrase. These seed patterns

¹ <http://www.ark.cs.cmu.edu/TweetNLP/>

are chosen manually based on our observations with the data. Although we observe that most of the time the noun phrases found in these patterns do indicate user interests, there are also cases when the noun phrases are not related to user interests, e.g. “life.” To ensure high precision, we then focus on the more frequent noun phrases. All extracted noun phrases are ranked based on their numbers of appearances in the whole corpus and the top-100 ranked phrases are selected as our seed interest tags. We annotate all occurrences of these seed interest tags with “B-INT” and “I-INT” in the biographies we have, regardless of whether they co-occur with our seed patterns.

4.2 Sequence Labeling using CRF

Both lexical and POS tag features are used in the CRF model. We do not use syntactic and dependency features as Twitter biographies are usually short and many of them are not sentences but phrases. As our training data is generated using seed interest tags, to avoid over-fitting, we only use contextual features, i.e. features extracted from the surrounding tokens for each position. The feature set we use is shown in Table 3. Different combinations of word features and POS tag features are used in our experiments. These features are empirically selected to get the optimal performance on our test data.

Table 3. Features we use for the CRF model

Feature	Description
Word Features	w_{i-1} w_{i+1}
Bigrams	$w_{i-2} + w_{i-1}$ $w_{i+1} + w_{i+2}$
PosTag	$POS_i + POS_{i+1}$ $POS_{i-1} + POS_i$
PosTag+Word	$POS_i + w_{i+1}$ $w_{i-1} + POS_i$

5 Experiments on Interest Tag Extraction

In this section, we present the empirical evaluation of the CRF-based method for interest tag extraction.

5.1 Data

We use a collection of 2,678,436 Twitter users’ biographies, which are all written in English. We preprocess these biographies by splitting sentences, tokenizing text and removing all punctuation marks and URLs.

For evaluation purpose, we created an annotated data set as follows. Two graduate students were recruited as human annotators. The annotators were asked to choose words or phrases that describe a user’s interest as interest tags. After discussion among the annotators, an annotation guideline about ambiguous cases was created and the annotators went through 500 randomly sampled biographies independently based on the

guideline. Out of the 1190 sentences from the 500 biographies, only 10 sentences have different annotations. We discard these 10 sentences and use the remaining sentences as ground truth in the following experiments.

5.2 Experiment Setup

Creation of training data: As we have discussed in Section 4, to create our noisy training data, we first obtain a set of seed interest tags and then use them to obtain a set of positive training sentences, which each contain at least one of the seed interest tags. We also randomly select sentences that do not contain any seed interest tag as our negative training data. After experimenting with different ratios of the positive and negative training sentences, we find that a ratio of 10:1 (positive to negative) gives the best performance.

Using CRF: We use CRF++², which is a C++ implementation of CRF. We use the default parameters of this implementation as our preliminary experiment shows that the performance of our model does not change much under different parameter settings.

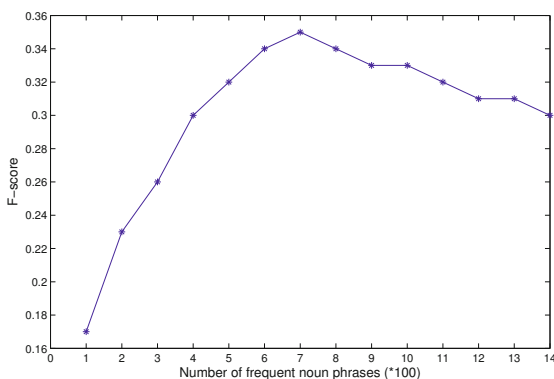


Fig. 2. F-score of baseline method over different number of frequent noun phrase

Baseline: As our task has not been studied before, there is no obvious baseline to compare against. We use two naive baselines for comparison. The first baseline, which is denoted by *Seed*, uses the seed pattern to extract interest tags directly. The second baseline first extracts all noun phrases from user biographies and then rank them by their numbers of appearances in the whole corpus. The top frequent noun phrases are selected and labeled as interest tags. As the performance of this baseline depends on the number of top frequent noun phrases we choose, we first conduct a preliminary experiments to choose the optimal number. Fig. 2 shows the performance in terms of F-score of using different numbers of top frequent noun phrases. The optimal value is 700. This method is denoted by *BL-700*.

² <http://crfpp.googlecode.com/svn/trunk/doc/index.html>

5.3 Results

Table 4 shows the performance of the trained CRF model and the baselines. While CRF have a comparable performance to *Seed*, it gets a much higher recall value. This shows that our seed patterns can only cover a small proportion of interest tags but a CRF classifier based on pseudo-labelled data generated by using seed patterns can extract more interest tags. We can also see that CRF outperforms *BL-700* substantially in all metrics, which shows the benefit of using supervised machine learning to perform the interest tag extraction task.

Table 4. Extraction performance of CRF and the baseline

	CRF	Seed	BL-700
Precision	0.91	0.92	0.22
Recall	0.65	0.03	0.76
F-score	0.76	0.06	0.35

Table 5 displays the top-10 frequent interest tags identified by CRF and BL-700. We can see that all interest tags extracted by CRF are meaningful words or phrases about users’ interests, but many words found by BL-700 are not describing interests.

Table 5. Top 10 interest tags extracted by our method and BL-700

CRF	BL-700
music	life
food	instagram
twitter	love
travel	music
coffee	god
tv	world
web	girl
internet	people
beer	everything
social media	things

5.4 Frequent Patterns

In this section, we show the frequently used triggers that Twitter users tend to use to indicate their interests. We first extract the adjacent n -grams right before and after each interest tag extracted by our method in the same sentence. Table 6 shows the unigram patterns around interest tags in Twitter users’ biographies. Words that can directly show interests are highlighted in bold font. We can see that Twitter users often use words such as “fanatic,” “fan,” “lover” and similar words to show their interests. In the list of unigram patterns right before interest tag, only “love” is a strong indicator. The other words are actually adjective words which are often used combined with words in the second

Table 6. N-gram patterns

Unigram		Bigram		Trigram	
Right Before	Right After	Right Before	Right After	Right Before	Right After
love	fanatic	I love	is my	to type with	is my life
Avid	fan	Subtly charming	glove on	I love one	happens for a
Wannabe	lover	Infuriatingly humble	is life	have a good	are my life
Extreme	enthusiast	type with	happens for	I am a	will be okay

column. For example, a user who is interested in football may describe himself as “extreme football fan” in his biography. Table 6 also shows patterns formed by bigrams and trigrams. We can see some clear patterns strongly indicating people’s interests or preferences, such as “I love,” “I like,” “in love with” and “is my life.”

6 Interest Tags and Tweets

One natural question to ask is what is the relationship between a user’s interest tags in biography and her tweet content. In other words, are the interest tags extracted from biographies reflected in users’ tweets? To study this problem, we treat interest tags as class labels and utilize tf-idf values of unigrams as features. For a given interest tag, we look at each user and predict whether or not this interest tag is relevant to her based on her tweets.

Tweets published by the users between September 2012 and August 2013 are collected. All conversational tweets that starts with “@”+username are removed. In each tweet, user names, retweet symbols, URLs and hashtags are also removed.

We extract the top-20 frequent interest tags and train a Logistic Regression classifier and an SVM classifier for each tag. The average accuracies based on 5-fold cross

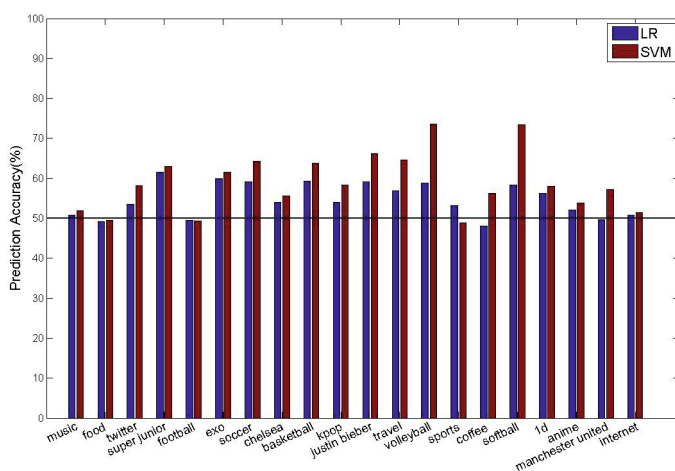


Fig. 3. Average accuracy for the top-20 tags.

validation are shown in Figure 3. The horizontal black line is the accuracy of random guess. We can see that the accuracy of most interest tags are just comparable to that of random guess and for some tags the performance is even worse than random guess. SVM classifier works better than LR for most tags. Both classifiers perform well in predicting tags such as “basketball”, “soccer,” “volleyball” and ”softball”, which are related to sports. This suggests that users with interest tags related to sports are more likely to tweet about things related to their interest tags. However, classifiers of those general interest tags, like “music”, “sports” and “Internet”, have very low accuracies, which indicates that these tags are harder to predict. One possible reason is that these tags are too general to be predicted.

We also consider another way to compare the interest tags extracted from biographies and users’ tweets. Here we try to extract tags from tweets and compare them with those extracted from biographies. Tf-idf ranking, a method that has been used to generate personalized user tags [21], is used in this task. In this method, all posted tweets of a user are grouped together into a single document and then tf-idf scores of terms in the document are calculated. The top- N highly scored words are extracted as user tags. We apply tf-idf ranking on our data set and compare the generated tags with the interest tags extracted from users’ biographies. For each user in the complete dataset, if we treat the interest tags from biographies as ground truth, we can calculate the recall of the tags extracted from tweets by tf-idf ranking. The results are shown in Table 7. We can see that the average recall of top-20 tags can only reach 7.2%, which is very low.

The above two experiments indicate that interest tags extracted from users’ biographies are not necessarily reflected in a users’ posted tweets. This suggests that using users’ tweets alone may not be sufficient and interest tags extracted from biographies may provide supplementary information of a user.

Table 7. Recall of tags from tweets.

Number of selected phrases	Recall of TF-IDF
top-5	0.045
top-10	0.057
top-15	0.066
top-20	0.072

7 Conclusion

In this work, we study the interest tags hidden in user biographies in Twitter. We first design a strategy based on a set of seed syntactic patterns to get noisy labeled training data. Then a sequential labeling algorithm CRF is trained based on this training data set. Our experiments show that the trained model gets very good performance. We also study the popular expressions people use to indicate their interests in biographies. We discover that tweet content may not reliably reflect users’ interest tags in biographies. Interest tags extracted from biographies provide supplementary information for users.

In the follow-up work, we are going to improve our sequential labelling model by introducing background knowledge to it, which could solve the low recall problem of

the current model. Demographic and professional information are also valuable and extraction of such information will be studied in the future. Combining information in biographies and social networks is also an interesting topic that will be explored in our future work.

Acknowledgements. This research is partially supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).

References

1. Benson, E., Haghghi, A., Barzilay, R.: Event discovery in social media feeds. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, pp. 389–398. Association for Computational Linguistics, Stroudsburg (2011)
2. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1301–1309. Association for Computational Linguistics, Stroudsburg (2011)
3. Counts, S., Stecher, K.B.: Self-presentation of personality during online profile creation. In: Proceedings of International AAAI Conference on Weblogs and Social Media, pp. 191–194. The AAAI Press, Dublin (2012)
4. Dong, W., Qiu, M., Zhu, F.: Who am i on Twitter?: A cross-country comparison. In: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion, pp. 253–254 (2014)
5. Filippova, K.: User demographics and language in an implicit social network. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1478–1488. Association for Computational Linguistics, Stroudsburg (2012)
6. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for Twitter: annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, pp. 42–47. Association for Computational Linguistics, Stroudsburg (2011)
7. Hagger-Johnson, G., Egan, V., Stillwell, D.: Are social networking profiles reliable indicators of sensational interests? *Journal of Research in Personality* 45(1), 71–76 (2011)
8. Imran, M., Elbassouni, S., Castillo, C., Diaz, F., Meier, P.: Practical extraction of disaster-relevant information from social media. In: Proceedings of the 22nd International Conference on World Wide Web Companion, pp. 1021–1024. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2013)
9. Lampe, C.A., Ellison, N., Steinfield, C.: A familiar face(book): Profile elements as signals in an online social network. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 435–444. ACM, New York (2007)
10. Liu, Z., Chen, X., Sun, M.: Mining the interests of Chinese microbloggers via keyword extraction. *Frontiers of Computer Science*, 76–87 (2012)
11. Marwick, A.E., et al.: I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society* 13(1), 114–133 (2011)

12. Mukherjee, A., Liu, B.: Improving gender classification of blog authors. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 207–217. Association for Computational Linguistics, Stroudsburg (2010)
13. Nguyen, D., Smith, N.A., Rosé, C.P.: Author age prediction from text using linear regression. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 115–123. Association for Computational Linguistics, Stroudsburg (2011)
14. Nosko, A., Wood, E., Zivcakova, L., Molema, S., De Pasquale, D., Archer, K.: Disclosure and use of privacy settings in Facebook profiles: evaluating the impact of media context and gender. *Social Networking*, 1–8 (2013)
15. Pennacchiotti, M., Popescu, A.M.: Democrats, republicans and Starbucks aficionados: user classification in Twitter. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 430–438. ACM, New York (2011)
16. Qiu, L., Lin, H., Leung, A.K.Y.: Cultural differences and switching of in-group sharing behavior between an American (Facebook) and a Chinese (Renren) social networking site. *Journal of Cross-Cultural Psychology* 44(1), 106–121 (2013)
17. Ritter, A., Clark, S., Mausam, E., Etzioni, O.: Named entity recognition in tweets: an experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1524–1534. Association for Computational Linguistics, Stroudsburg (2011)
18. Ritter, A., Mausam, E.O., Clark, S.: Open domain event extraction from Twitter. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1104–1112. ACM, New York (2012)
19. Rosenthal, S., McKeown, K.: Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 763–772. Association for Computational Linguistics, Stroudsburg (2011)
20. Roshchina, A., Cardiff, J., Rosso, P.: User profile construction in the twin personality-based recommender system. In: Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology, pp. 73–79. Asian Federation of Natural Language Processing, Chiang Mai (2011)
21. Wu, W., Zhang, B., Ostendorf, M.: Automatic generation of personalized annotation tags for twitter users. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 689–692. Association for Computational Linguistics, Stroudsburg (2010)
22. Yang, H., Li, Y.: Identifying user needs from social media. IBM Research report (2013)