Towards User Personality Profiling from Multiple Social Networks

Kseniya Buraya[1] and **Aleksandr Farseev**[2] and **Andrey Filchenkov**[3] and **Tat-Seng Chua**[4]
[1,3]ITMO University
49 Kronverksky Pr., St. Petersburg, 197101, Russia, +78122329704
[2,4]National University of Singapore
21 Lower Kent Ridge Rd, Singapore, 119077, +6565166666
Email: ksburaya@corp.ifmo.ru, farseev@u.nus.edu, afilchenkov@corp.ifmo.ru, dcscts@nus.edu.sg

Abstract

The exponential growth of online social networks has inspired us to tackle the problem of individual user attributes inference from the Big Data perspective. It is well known that various social media networks exhibit different aspects of user interactions, and thus represent users from diverse points of view. In this preliminary study, we make the first step towards solving the significant problem of personality profiling from multiple social networks. Specifically, we tackle the task of relationship prediction, which is closely related to our desired problem. Experimental results show that the incorporation of multi-source data helps to achieve better prediction performance as compared to single-source baselines.

User profiling plays an increasingly important role in many application domains (Farseev, Samborskii, and Chua 2016). One of the critical components of user profiling is personality profiling (Pennebaker, Mehl, and Niederhoffer 2003), which seeks to identify one's mental and emotional characteristics. Knowing these personal attributes can help to understand reasons behind one's behaviour (Pennebaker, Mehl, and Niederhoffer 2003), select suitable individuals for particular tasks (Song et al. 2015), and motivate people to undertake new challenges in their life. Up to now, there have been several research attempts towards personality profiling. For example, some research groups have investigated this problem from the social science point of view (Pennebaker, Mehl, and Niederhoffer 2003). However, most of these works are descriptive in nature and rely on manual data collection procedures, which explains the absence of large-scale research in the field. With the recent growth of the Web, personality profiling can be approached by taking advantage of the abundance of data from online social networks. For example, such data has been utilized by several studies and evaluations devoted to automatic personality profiling, such as TwiSty (Verhoeven, Daelemans, and Plank 2016) or PAN (Rangel et al. 2015). Even though these studies made a significant progress towards automatic personality profiling, most of them were carried out on data from a single source (i.e. Twitter) or of a single modality (i.e. Text). Such personality profiling may lead to a sub-optimal performance (Farseev and Chua 2017). Taking into account that most social networks users use more than one social network in their daily life (Farseev et al. 2015a), it is reasonable to utilize multiple data sources and modalities to solve personality profiling task.

There are several personality categorization schemes adopted by the research community. One of the most widely embraced typologies is called Myers-Briggs Type Indicator (MBTI), that was proposed by Mayer and Briggs in 1985 and based on Carl Jung's theory. The typology is designed to *exhibit psychological preferences on how people perceive the world around them* and distinguishes 16 personality types. Meanwhile, it was also discovered that *social media services exceedingly affect and reflect the way their users communicate with the world and among themselves* (Kaplan and Haenlein 2010). Based on these observations, it follows that MBTI categorization schema naturally fits social media research. Further, according to the previous studies (Farseev et al. 2015b; Farseev and Chua 2017) and our findings, social media users reveal their personal attributes differently in different social media platforms. For example, they may post photos in photo-sharing services, such as Instagram, or perform check-ins in location-based social networks, such as Foursquare. All this data describes users from the 360° view and, thus, plays an essential role in social media-based personality profiling.

However, personality profiling from multiple social networks is associated with the following challenges:

- **Cross-source user identification.** Often, it is not possible to identify multiple social networks accounts that belong to the same person, while some users use a limited number of social networks.
- **Ground-truth collection.** Not all online resources with MBTI information about their users are approved by psychologists, while only a limited number of social networks posts is equipped with the references to trusted MBTI profiling resources.
- **Temporal changes of users' personality.** Users' personality trends vary over time under the influence of different life aspects and external factors, which requires additional consideration during the data modeling process.
- **Data source fusion.** Effective fusion of multi-view data from different sources in one model is a challenging problem (Song et al. 2015).

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Inspired by the above challenges, in this preliminary study we formulate our research question as: **is it possible to increase the performance of personality profiling by incorporating data from multiple social networks?**

Due to the absence of multi-source personality profiling datasets, in this study, we utilize the largest available multi-source cross-region dataset NUS-MSS (Farseev et al. 2015b), which includes multi-modal data from three social networks (Twitter, Foursquare, and Instagram) and the ground truth records regarding users' relationship status. The data is provided for three geographical regions, namely Singapore, New York, and London.

Robins et al. (2002) postulated, that relationship status is closely related to human' personality. Indeed, one's personality is often affected by his/her current relationship status, and inversely, one's relationship status often depends on life partners' personality types match. Moreover, similarly to personality type, the relationship status attribute can also be considered a dynamic personal attribute, since it often changes over time. All the above suggest that relationship status is closely associated with one's personality, which inspired us to select it for evaluation.

To perform quantitative evaluation, we divided NUS-MSS users into "single", and "not single" groups. We then applied feature selection and evaluated the classification performance in terms of average accuracy in three geographical regions, namely Singapore, New York, and London.

The evaluation results are presented in Table 1. In this study, we only investigated the early fusion of different combinations of data sources, which is concatenation of sources' feature vectors in one vector before model training. From Table 1, it is clear that the early fusion of multi-source data helps to improve classification performance by more than 17% in some cases. It thus answers our research question positively and suggests the usage of multi-source data in our further works on personality profiling.

However, it is noted that early fusion of three data sources fails to achieve the best performance for all three geographical regions. This seems to indicate that early fusion approach is unable to fully exploit the richness and diversity of all data sources efficiently. This observation is consistent with previous findings (Song et al. 2015; Farseev and Chua 2017) and inspires us to explore other data fusion strategies, including late fusion, in our future work. At last, Robins et al. (2002) has shown that changes in relationship status are often correlated with changes in one's personality. In future work, we thus also plan to utilize relationship status information for model regularization, which could potentially boost personality profiling performance.

On personality-related data collection: To perform further experiments on personality profiling, it is important to obtain data from multiple social networks and the corresponding personality-related ground truth. To do so, we plan to enrich existing datasets on personality profiling by data from multiple social networks, which can be accomplished by monitoring users' cross-posting activity (Farseev et al. 2015b; Farseev and Chua 2017). Our initial experiments reveal that over 66% of Twitter users with personality ground truth can be mapped to their Instagram or Foursquare accounts.

Table 1: Relationship status prediction evaluation results

	Singapore	New York	London
Twitter (Twi)	0.688	0.580	0.637
Instagram (Inst)	0.677	0.540	0.613
Foursquare (Fs)	0.666	0.800	0.590
Twi, Inst	**0.865**	**0.851**	0.655
Twi, Fs	0.790	0.714	**0.808**
Inst, Fs	0.780	0.710	0.714
Twi, Inst, Fs	*0.780*	*0.714*	*0.714*

Acknowledgements

NExT research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC@SG Funding Initiative. This research is partially supported by the Government of Russian Federation, Grant 074-U01 and JetBrains Research.

References

Farseev, A., and Chua, T.-S. 2017. Tweetfit: Fusing multiple social media and sensor data for wellness profile learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI.

Farseev, A.; Kotkov, D.; Semenov, A.; Veijalainen, J.; and Chua, T.-S. 2015a. Cross-social network collaborative recommendation. In *Proceedings of the ACM Web Science Conference*, 38. ACM.

Farseev, A.; Nie, L.; Akbari, M.; and Chua, T.-S. 2015b. Harvesting multiple sources for user profile learning: a big data study. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 235–242. ACM.

Farseev, A.; Samborskii, I.; and Chua, T.-S. 2016. bbridge: A big data platform for social multimedia analytics. In *Proceedings of the 2016 ACM on Multimedia Conference*, 759–761. ACM.

Kaplan, A. M., and Haenlein, M. 2010. Users of the world, unite! the challenges and opportunities of social media. *Business horizons* 53(1):59–68.

Myers, I. B.; McCaulley, M. H.; and Most, R. 1985. *Manual, a guide to the development and use of the Myers-Briggs type indicator*. Consulting Psychologists Press.

Pennebaker, J. W.; Mehl, M. R.; and Niederhoffer, K. G. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54(1):547–577.

Rangel, F.; Rosso, P.; Potthast, M.; Stein, B.; and Daelemans, W. 2015. Overview of the 3rd author profiling task at pan 2015. In *CLEF*.

Robins, R. W.; Caspi, A.; and Moffitt, T. E. 2002. It's not just who you're with, it's who you are: Personality and relationship experiences across multiple relationships. *Journal of personality* 70(6):925–964.

Song, X.; Nie, L.; Zhang, L.; Akbari, M.; and Chua, T.-S. 2015. Multiple social network learning and its application in volunteerism tendency prediction. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 213–222. ACM.

Verhoeven, B.; Daelemans, W.; and Plank, B. 2016. Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*.