# A Whole New Ball Game: Harvesting Game Data for Player Profiling

**Ivan Samborskii,[1] Aleksandr Farseev,[2] Andrey Filchenkov,[2] Tat-Seng Chua[1]**
[1] National University of Singapore
21 Lower Kent Ridge Rd, Singapore, 119077, +6565166666
[2] ITMO University
49 Kronverksky Pr., St. Peterburg, 197101, +78122329704
i.samborskii@u.nus.edu, farseev@gmail.com, afilchekov@corp.ifmo.ru, dcscts@nus.edu.sg

## Abstract

Nowadays, video games play a very important role in human life and no longer purely associated with escapism or entertainment. In fact, gaming has become an essential part of our daily routines, which give rise to the exponential growth of various online game platforms. By participating in such platforms, individuals generate a multitude of game data points, which, for example, can be further used for automatic user profiling and recommendation applications. However, the literature on automatic learning from the game data is relatively sparse, which had inspired us to tackle the problem of player profiling in this first preliminary study. Specifically, in this work, we approach the task of player gender prediction based on various types of game data. Our initial experimental results inspire further research on user profiling in the game domain.

During the past decade, the game industry has faced a surge of data, which includes tons of behavioural information (Bohannon 2010), including, for example, game preferences, various statistics, and in-game interaction sessions. Such information is quite heterogeneous most of the time, which complicates its processing. To structure such information, the so-called "user profiling" techniques are often used.

In this work, we define user profiling, specifically player profiling, as the process of automatic player attribute (e.g., demographic) inference from player-generated data. Knowing player attributes is beneficial for various applications. For example, in the public sector, they can be employed for early game addiction tendency identification (Bauckhage, Drachen, and Sifa 2015). At the same time, the predicted player attributes can be adopted to enhance the accuracy of the game recommendation services, game design improvement, and game experience personalization.

Up to now, there have been several research attempts performed towards data analysis in the game research domain. For example, social scientists have examined the problem from the qualitative perspective (Yee 2006). However, most such works are descriptive in nature and rely on manually collected data, which explains the absence of large-scale data sets in the field. In the course of time, the research attempts on player profiling have been re-focused towards predictive analysis. For example, Yang et al. (2018) inferred players' game experiences based on different human signals recorded by various sensors. Even though the study has shown significant progress towards automatic player profiling, it, again, has been carried out based on a small-scale dataset collected in a "sterile" lab environment, which does not allow for making large-scale conclusions. Nevertheless, the tremendous amounts of data in the game domain open the opportunity to perform player profiling from the Big Data perspective and this study is the first preliminary attempt towards such an exciting research direction.

Despite a wide availability of the data, player profiling is associated with the following challenges:

- **Ground truth collection**. Most players prefer to stay anonymous in the virtual world, as well as behave differently as compared to their social behaviour in the real world. For example, often male players choose female game personage, thus ground truth records must be obtained from other data sources, such as conventional social networks.
- **Cross-source identification**. In order to obtain user-related ground truth records (e.g., gender labels), one needs to hold information about linkages to other social network accounts, such as Facebook, Instagram, etc.
- **Longitudinal variations**. Players may change their behaviour over time by, for example, experiencing evolving game scenarios (Bauckhage, Drachen, and Sifa 2015), etc. Such behavioural fluctuations require to be handled properly during the data modelling process.
- **Data fusion**. Effective fusion of the multi-modal game data (e.g., player's chat, in-/out-game activities, etc.) in one model in a mutually-consistent fashion is a challenging research problem (Farseev et al. 2015).

Inspired by the challenges above, we formulate our research question as follows: **is it possible to learn from multi-view temporal game experience data for player profiling?**

Due to the absence of publicly-available game datasets, in this study, the dataset was collected from scratch via monitoring of the publicly available game accounts. To do so, we took advantage of the public data available at Player.me game social network. One of the network's features is the explicit linkage between social accounts and game accounts belonging to the same individuals. Leveraging on such linkages, we have collected the dataset, which contains data from multiple social accounts with respect to the same players. Steam game platform was chosen as the source of players' game activity. The data has been collected during the time interval between 19 January 2017 and 3 April 2018, while, simultaneously, the gender ground truth records were gathered from players' Facebook accounts.

Based on the collected data we have extracted six different data representations: (1) *game statistics* features: (a) $u_i \rightarrow \{q_1, ..., q_m\}$, where $q_j$ is the number of times user $u_i$ played game $j$; and (b) $u_i \rightarrow \{qt_1, ..., qt_m\}$, where $qt_j$ is the time in minutes that user $u_i$ played game $j$; (2) *game genre* features: (a) $u_i \rightarrow \{g_1, ..., g_k\}$, where $g_j$ is the number of times user $u_i$ played games of $j$'s genre (e.g., Action, RPG, etc.); and (b) $u_i \rightarrow \{gt_1, ..., gt_k\}$, where $gt_j$ is the time in minutes that user $u_i$ played games of genre $j$; and (3) *game category* features: (a) $u_i \rightarrow \{c_1, ..., c_p\}$, where $c_j$ is the number of times user $u_i$ played games of $j$'s category (e.g., Co-op, Mods, etc.); and (b) $u_i \rightarrow \{ct_1, ..., ct_p\}$, where $ct_j$ is the time in minutes that user $u_i$ played games of category $j$.

Based on our literature survey, the related studies often suggest the use of data fusion approaches for multi-view learning, which was found quite effective for solving the data integration problem for player profiling (Farseev and Chua 2017). Specifically, the so-called early and late fusion methods were extensively adopted. The strategy of early fusion is to concatenate all data representation vectors into one common vector prior to learning from the data, while the late fusion suggests the combination of the pre-trained on different data representations classifiers: $f(x) = \arg\max_j \left(\sum_i (w_i \cdot \text{confidence}_{ij}(x))\right)$, where $i$ and $j$ are a classifier and class indices, respectively.

In this study, we have adopted the late fusion strategy, described in Farseev et al. (2015), where individually-trained Random Forests are linearly combined together in an ensemble and the combinations weights are optimized by Stochastic Hill Climbing with Random Restart. Other state-of-the-art time-tested models that have shown their effectiveness for many tasks ($k$NN, Naïve Bayes, SVM, Decision Tree, Random Forest (RF), and Random Forest-based Early Fusion) were included as evaluation baselines.

To assess our approach against the baselines, we divided the dataset users into "male" and "female" groups. Particularly, 984 male and 711 female player accounts from Steam game platform were included. For all chosen models, we applied feature selection as described in (Samborskii et al. 2016) and then evaluated the classification performance in terms of F-score via 10-Fold Cross-Validation.

The evaluation results are presented in the Table 1. It can be seen that although Early Fusion performs better than other baselines, it is not able to cope with large feature spaces effectively, which explains the superiority of Late Fusion approach by over than 11% in some cases. Such an observation positively answers our research question and inspires further studies on multi-view temporal learning from data in the game domain.

We would also like to note that our obtained results could be further improved. To do so, in our future studies we will: (1) make additional efforts towards diversifying feature spaces (e.g., include spatial data), which is expected to further improve the classification performance; (2) utilize the data temporality at the learning stage, which is expected to handle higher-order temporal relationships as compared to vanilla feature engineering approaches; (3) enrich existing dataset by the data from conventional social networks, such as Twitter, Instagram, and Pinterest, which will help us to learn player profiles from a 360° perspective (Farseev et al. 2015); and (4) extend the existing ground truth records to facilitate other player profiling tasks, such as age group and personality identification.

Table 1: Gender prediction evaluation results

| | Data repr. | Classifier | F-score |
|---|---|---|---|
| Single-View | Games | RF (55 trees) | 0.580 |
| | Games (time) | RF (111 trees) | 0.588 |
| | Genres | RF (66 trees) | 0.555 |
| | Genres (time) | RF (139 trees) | 0.557 |
| | Categories | RF (107 trees) | 0.557 |
| | Categ. (time) | RF (158 trees) | 0.557 |
| Multi-View | Early Fusion | RF (98 trees) | 0.606 |
| | Late Fusion ($w_q = 0.58, w_{qt} = 0.31, w_g = 0.10, w_{gt} = 0.15, w_c = 0.25, w_{ct} = 0.10$) | | $\mathbf{0.668}$ |

## Acknowledgements

NExT research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC @SG Funding Initiative. This research is supported by the Government of Russian Federation, Grant 08-08.

## References

Bauckhage, C.; Drachen, A.; and Sifa, R. 2015. Clustering game behavior data. *IEEE Transactions on Computational Intelligence and AI in Games* 7(3):266–278.

Bohannon, J. 2010. Game-miners grapple with massive data. *Science* 330.

Farseev, A., and Chua, T.-S. 2017. Tweet can be fit: Integrating data from wearable sensors and multiple social networks for wellness profile learning. *ACM Transactions on Information Systems (TOIS)* 35(4):42.

Farseev, A.; Nie, L.; Akbari, M.; and Chua, T.-S. 2015. 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.

Samborskii, I.; Filchenkov, A.; Korneev, G.; and Farseev, A. 2016. Person, organization, or personage: Towards user account type prediction in microblogs.

Yang, W.; Rifqi, M.; Marsala, C.; and Pinna, A. 2018. Towards better understanding of player's game experience. In *ACM ICMR 2018*, 442–449. ACM.

Yee, N. 2006. Motivations for play in online games. *CyberPsychology & behavior* 9(6):772–775.