Analysis of the Current Situation and Trends in Player Modelling Based on Data

Guitan Zhang

{zhangg58@mcmaster.ca}

Faculty of Science, McMaster University, Ontario, L8S 4L8, Canada

Abstract. Currently, player customization is gaining popularity, leading to increased attention on the field of player modeling. This paper takes a focus on data, analyzes research trends in player modelling, and explores the commercial and technological reasons behind it. This article uses pie charts to visualize the data and reveals the trends in types, including feature, label, source of data, and way of sampling in this field after 2020. This paper argues that current research in player modelling can be categorized into three types: research predicting objective labels that obtain data from the internet, research predicting objective labels that require obtaining data through experiments independently, and clustering experiments that can obtain data across fields since they do not require labels. This paper can provide a basic reference for scholars conducting subsequent research in related fields.

Keywords: Data analysis, player modelling, game-play-based

1 Introduction

Customized content refers to content that is different for each player, chosen by a decision made by either the player themselves or the game program, within the same game. In 2008, Left 4 Dead used a novel AI director, with the purpose to adjust difficulty and ensure gammin experience for players. After that, Minecraft offers free version selection and mod interfaces, allowing players to create and install customized content.

As a turning point, Valve launched the Workshop feature on Steam, providing many games with the possibility of building communities where players can create, share, and install customized content easily. Since then, many games have caught this opportunity, and succeed on extending their lifespan due to the large amount of customized content created by players in Workshop, and based on that, the question of whether to open access to Steam Workshop becomes a question need to be decided for many game developers.

As early as 2012, Yannakakis noticed that the application of artificial intelligence in the gaming field was maturing; he reviewed the interaction between AI and games and categorized the application of artificial intelligence in the gaming field into four key directions: Player

ICIAAI 2026, May 22-24, Singapore, Singapore
Copyright © 2026 EAI
DOI 10.4108/eai.22-5-2026.2365138

Experience Modeling, Procedural Content Generation, Massive-Scale Game Data Mining, and NPC AI[1]. In 2023, Yannakakis introduced the methods, tools, and definitions of affective computing through the concept of affective game loops, introduced many papers in this field, reviewed commercial examples of affective computing in the game field, introduced some current dilemmas in the field, and he is optimistic about research in the fields of general models, computer vision, and large language models[2]. Meanwhile, Paraschos, from an application perspective, integrated research on dynamic difficulty adjustment and procedural content generation from 2005 to 2021, reviewed the prospects for the integration of this field with serious games, and also affirmed the help that player modeling and personalized games provide to players, especially their contribution to patient rehabilitation[3]. Subsequently, in 2025, Yannakakis reviewed the field of player modeling. He described types of models, inputs, and outputs in this area, mentioned the contributions of player modeling to game design, and also discussed the many challenges in data, technology, and theory within this field[4].

Player modeling can be understood as a method of modeling players' cognition, behavior, and emotional state based on data obtained from human players' interactions with games. Building upon the definition proposed by Yannakakis [1,4], the purpose of this paper is to identify new trends in paper data after 2020, based on a comparison with research directions before 2020. By collecting research in this field, this paper focuses on features, label values, data sources, and sampling methods, analyzing them over time and providing possible reasons for these changes.

2 Analysis of Data Source Distribution

This article uses Google Scholar's AI Labs feature to search for research related to player modeling and collects predictive experiments related to game players (such as churn rate, next action, classification, game experience, etc.). Several papers on World of Warcraft appeared, but this article only includes one.

This paper categorizes the experiment into five aspects based on manual identification: "After 2020" refers to whether the paper associated with the experiment was published after 2020, based on the time displayed by Google Scholar. “Feature” refers to a feature used in the study. GPB (game-play-based) means that the study involves the time series of telemetry game input specified by the researcher as the feature value. Data represents other features. “Label” refers to labels used in research. Subjective means that the labels are made based on the opinions of experimenters (votes, questionnaires, etc.). In the experiment, the experimenters are fully aware that they are inputting labels, and they can freely change their corresponding labels if they wish. Objective means that labels are made from objective data (physiological data, next steps, dropout rate, etc.). The experimenters are often unaware that they are inputting data, and even if they wish, they often cannot change their corresponding labels. Clustering refers to classification algorithms where there are no label values. This paper classifies them as a category of player modeling, since it is worth noting that after conducting classification experiments, experimenters often analyze the data of each group to model player authentication and behavior. "Data" refers to the types of data sources researchers use. Experiment means the data is collected by the researcher themselves, Company means the data comes from a company, and the same applies to Other Papers and Websites. "Sample" refers to the sampling method used in the study. Gameplay means the researcher is present when the experimenter inputs data, while Internet

means the researcher is not present, often indicating that the experiment was conducted by the experimenter across the internet. When the data source is not an experiment, the sample is always the internet.

Based on the above experimental steps, this paper selected 22 experiments from 19 papers for analysis, as shown in Table 1.

Table 1. Experiments and categories from 5 perspectives

Paper	After 2020	Feature	Label	Data	Sample
[5]	no	data	subjective	company	internet
[6]	yes	GPB	clustering	company	internet
[7]	yes	data	subjective	experiment	internet
[8]	yes	data	objective	website	internet
[9]	no	data	clustering	experiment	gameplay
[10]	yes	GPB	clustering	experiment	gameplay
[10]	yes	data	objective	experiment	gameplay
[11]	no	data	subjective	experiment	gameplay
[12]	no	data	subjective	experiment	gameplay
[13]	no	data	objective	experiment	internet
[13]	no	data	objective	experiment	internet
[14]	no	GPB	subjective	experiment	gameplay
[15]	no	data	subjective	experiment	gameplay
[16]	no	data	subjective	experiment	internet
[17]	no	data	clustering	company	internet
[17]	no	data	clustering	website	internet
[18]	yes	data	clustering	experiment	survey
[19]	no	data	clustering	other papers	internet
[20]	yes	GPB	objective	website	internet
[21]	yes	data	objective	company	internet
[22]	yes	GPB	objective	company	internet
[23]	yes	GPB	subjective	experiment	gameplay

Among them, the experiment of Jiang, S., Zhang, L., Xu, H., Huang, J., He, Q., Zhou, X., ... & Jiang, J. used data from the company as feature values and data from the website as label values [20], which is labeled as “website”, but it does not affect the conclusion.

2.1 AnaAnalysis based on the comparison before and after 2020

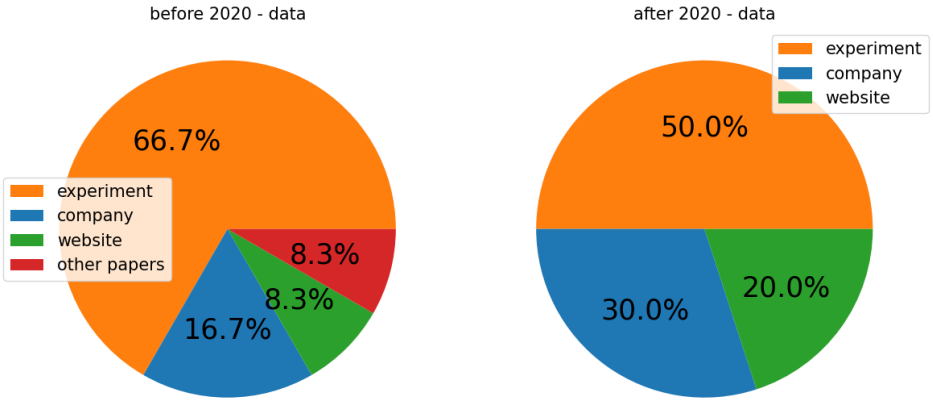

Fig. 1. Distribution of data before and after 2020(Photo/Picture credit: Original)

As can be seen from Figure 1, the proportion of research using data from companies and websites increased from 25% to 50%, which means that the cooperation between companies and research institutions is gradually increasing. From the perspective of companies, this may be due to the company's need to understand player preferences. In the research using websites as the source, Rismayanti used a dataset with player retention rate as the target for training [8], while Jiang, S., Zhang, L., Xu, H., Huang, J., He, Q., Zhou, X., ... & Jiang, J. extracted data as only label values from a large dataset on GitHub, with features data extracted from companies [20].

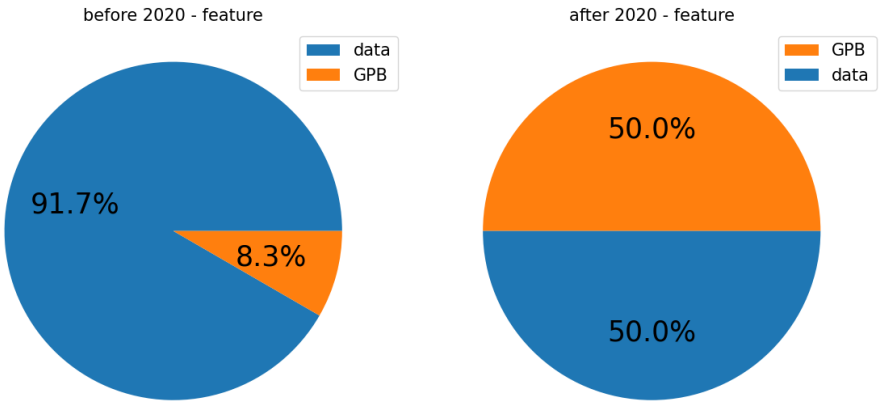

Fig. 2. Distribution of feature before and after 2020(Photo/Picture credit: Original)

As shown in Figure 2, the proportion of researchers using GPB as feature values surged from 8.3% to 50%. This phenomenon is the result of multiple factors: the field has been influenced by the technological boom of attention models for next token prediction. For example, in 2024, Wang, T., Honari-Jahromi, M., Katsarou, S., Mikheeva, O., Panagiotakopoulos, T., Asadi, S., & Smirnov, O. used the longformer (2020) studied by Beltagy, I., Peters, M. E., & Cohan, A., a type transformer model for long sequences [6,24]. Meanwhile, breakthroughs in data annatiting tool and data processing technologies have also contributed to this phenomenon. For

example, in 2021, Zhao, S., Xu, Y., Luo, Z., Tao, J., Li, S., Fan, C., & Pan, G. used the distributed coding of Le, Q., & Mikolov, T. (2014), which was originally a text processing technique [22,25]. In 2024, Fernandes, P. M., Lopes, M., & Prada, R. cited the research of Lopes, P., Yannakakis, G. N., & Liapis, A. (2017), which is a technique that allows experimenters to continuously label their game experiences, solving the problem that previously only player experience summaries after game completion could be obtained through questionnaires [6,26]. Finally, this phenomenon has also been influenced by company support: according to Table 1, in experiments that obtained feature data from companies after 2020, 75% used the time series of game input as features [6,20,21,22].

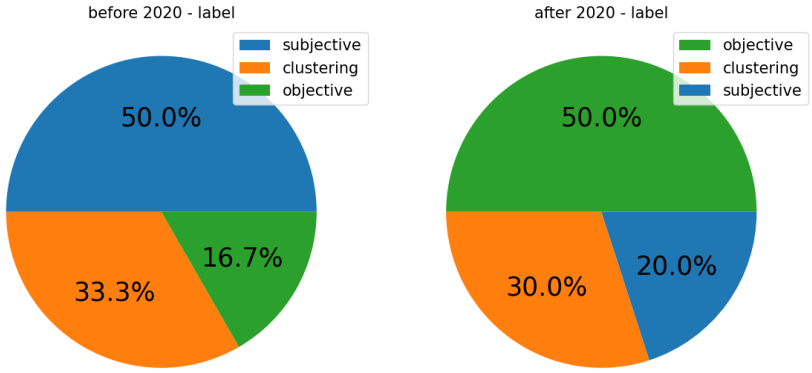

Fig. 3. Distribution of label before and after 2020(Photo/Picture credit: Original)

As shown in Figure 3, after 2020, the proportion of experiments using subjective labels decreased from 50% to 20%, while the proportion of experiments using objective labels increased from 16.7% to 50%. This is also due to the influence of companies and websites on the data: according to Figure 4, after 2020, 80% of the experiments using objective labels used data from companies or websites.

3. Discussion

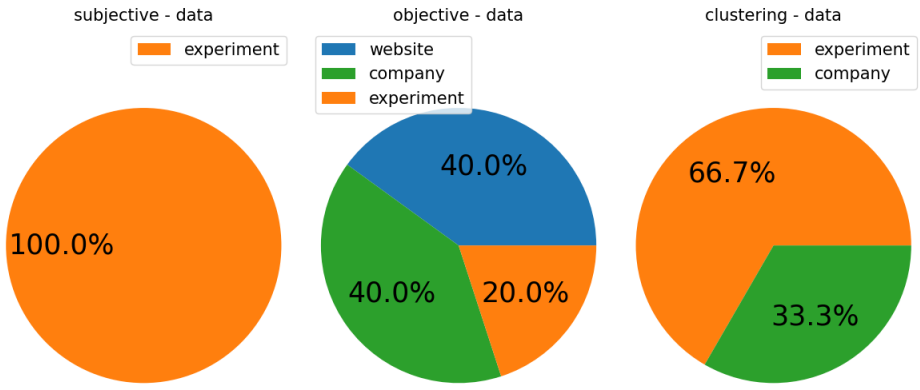

Fig. 4. Distribution of data for different labels after 2020(Photo/Picture credit: Original)

As shown in Figure 4, the collection of subjective labels is clearly absolutely dependent on the experimental data. This indicates that there are currently almost no publicly available datasets with subjective labels applicable to this field, forcing researchers to collect data manually. In contrast, the path to collecting subjective labels depends on asking experimenters for their opinions, and the clustering algorithm is completely independent of labels. This suggests that this phenomenon may be due to the relatively high difficulty, slow speed, and low universality of collecting subjective labels.

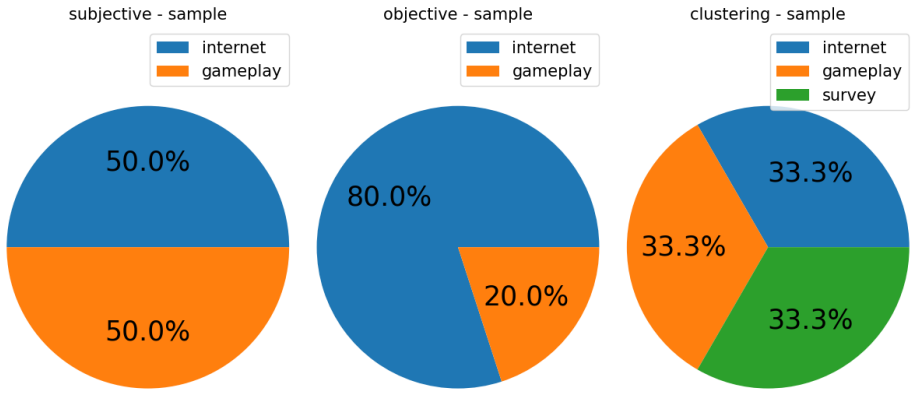

Fig. 5. Distribution of samples for different labels after 2020(Photo/Picture credit: Original)

As shown in Figure 5, subjective labels and clustering algorithms fail to reveal any bias towards a particular sampling method, while objective labels clearly rely on online sampling. This can be explained from the perspective of the company's data mining technology: current telemetry technology allows companies to mine various objective data as labels, from each player on servers, but if the company needs subjective data as labels, it needs to carefully design questionnaires, given the large number of participating users. Furthermore, distributing questionnaires online inevitably leads to many people not taking them seriously, so the company needs to perform proper data cleaning after collection to avoid data distortion caused by invalid data, and these operations themselves incur costs. More importantly, distributing questionnaires to users can lead to user dissatisfaction, thus impacting company operations.

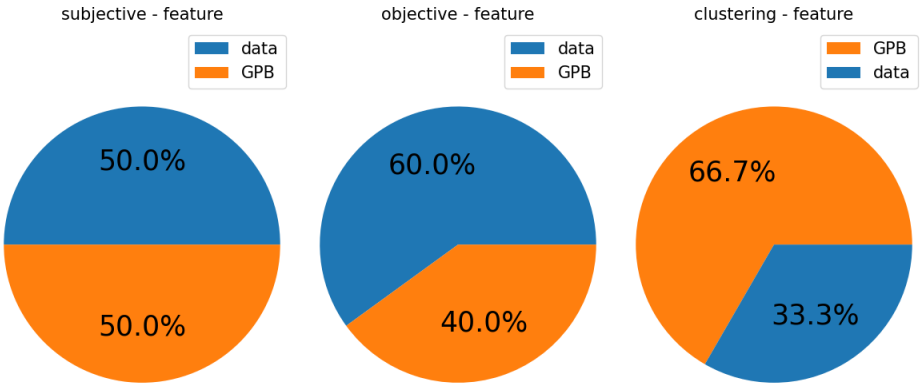

Fig. 6. Distribution of features for different labels after 2020(Photo/Picture credit: Original)

As can be seen from Figure 6, the feature distribution did not show a significant difference compared to other distributions (difference <16.7%). This may indicate that researchers, regardless of which type of feature they intend to collect, can consider collecting two labels and simultaneously attempting to conduct research in the direction of all categories of labels.

By focusing on clustering algorithms and analyzing Figures 4, 5, and 6, clustering algorithms show almost no obvious directional bias. This paper hypothesizes that this is because clustering algorithms do not rely on labels. Therefore, current research on clustering algorithms may have stemmed from researchers who, after collecting subjective or objective labels, found themselves unable to produce convincing results and thus chose clustering as their research focus.

4 Conclusion

When a researcher attempts to conduct predictive experiments, the first thing to consider, before thinking about the choice of model and technique, is how to collect data.

This paper summarizes the current trends and possible reasons by marking the proportion of different types of data in the papers, focusing on time (before 2020 vs. after 2020) and labels (subjective, objective, clustering) after 2020.

This paper argues that after 2020, the proportion of studies using GBP data as features has significantly increased, as has the proportion using objective data as labels, while the proportion using subjective data as labels has decreased. This is attributed to the increasing data support from companies and websites, new methods for time-series features brought about by attention models, and new breakthroughs in data annotating and processing techniques.

Meanwhile, this paper argues that three main directions are emerging in this field: research on predicting objective labels based on website and company data, research on predicting subjective labels based on data obtained through experiments, and clustering research that does not rely on labels.

This paper aims to provide a basic reference for researchers attempting to conduct research in this field.

This paper only categorizes data and predicts possible future directions, but does not engage in a deep discussion about the technologies used or the specific research process. Future research could delve deeper into existing literature from the perspective of technology application. Furthermore, the search in this paper fully relied on the AI Labs feature of Google Scholar; more rigorous and scientific search methods can be explored in the future.

References

[1] Yannakakis, G.N.: Game AI revisited. Proceedings of the Conference on Computing Frontiers. pp. 285–292 (2012)

[2] Yannakakis, G.N.; Melhart, D.: Affective game computing: A survey. Proceedings of the IEEE. 111(10), 1423–1444 (2023)

[3] Paraschos, P.D.; Koulouriotis, D.E.: Game difficulty adaptation and experience personalization: A literature review. International Journal of Human–Computer Interaction. 39(1), 1–22 (2023)

[4] Yannakakis, G.N.; Togelius, J.: Player modeling. In: Artificial Intelligence and Games. Springer, Cham, pp. 315–335 (2025)
[5] Blackburn, J.; Kwak, H.: STFU NOOB! Predicting crowdsourced decisions on toxic behavior in online games. Proceedings of the International World Wide Web Conference. pp. 877–888 (2014)
[6] Wang, T.; Honari-Jahromi, M.; Katsarou, S.; Mikheeva, O.; Panagiotakopoulos, T.; Asadi, S.; Smirnov, O.: Player2vec: A language modeling approach to understand player behavior in games. ArXiv preprint (2024)
[7] Quwaider, M.; Alabed, A.; Duwairi, R.: Shooter video games for personality prediction using five-factor model traits and machine learning. Simulation Modelling Practice and Theory. 122, 102665 (2023)
[8] Rismayanti, N.: Predicting online gaming behaviour using machine learning techniques. Indonesian Journal of Data and Science. 5(2), 133–143 (2024)
[9] Nogueira, P.; Aguiar, R.; Rodrigues, R.; Oliveira, E.; Nacke, L.: Fuzzy affective player models: A physiology-based hierarchical clustering method. Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. 10(1), 132–138 (2014)
[10] Fernandes, P.M.; Lopes, M.; Prada, R.: Generating game levels by defining player experiences. Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. 20(1), 179–188 (2024)
[11] Makantasis, K.; Liapis, A.; Yannakakis, G.N.: From pixels to affect: A study on games and player experience. Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 1–7 (2019)
[12] Thue, D.; Bulitko, V.; Spetch, M.; Romanuik, T.: A computational model of perceived agency in video games. Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. 7(1), 91–96 (2011)
[13] Weber, B.G.; Mateas, M.; Jhala, A.: Using data mining to model player experience. Proceedings of the FDG Workshop on Evaluating Player Experience in Games. pp. 1–8 (2011)
[14] Guckelsberger, C.; Salge, C.; Gow, J.; Cairns, P.: Predicting player experience without the player: An exploratory study. Proceedings of the ACM Symposium on Computer-Human Interaction in Play (CHI PLAY). pp. 305–315 (2017)
[15] Shaker, N.; Asteriadis, S.; Yannakakis, G.N.; Karpouzis, K.: Fusing visual and behavioral cues for modeling user experience in games. IEEE Transactions on Cybernetics. 43(6), 1519–1531 (2013)
[16] Pedersen, C.; Togelius, J.; Yannakakis, G.N.: Modeling player experience for content creation. IEEE Transactions on Computational Intelligence and AI in Games. 2(1), 54–67 (2010)
[17] Drachen, A.; Sifa, R.; Bauckhage, C.; Thurau, C.: Guns, swords and data: Clustering of player behavior in computer games in the wild. Proceedings of the IEEE Conference on Computational Intelligence and Games (CIG). pp. 163–170 (2012)
[18] Potard, C.; Henry, A.; Boudoukha, A.H.; Courtois, R.; Laurent, A.; Lignier, B.: Video game players' personality traits: An exploratory cluster approach to identifying gaming preferences. Psychology of Popular Media. 9(4), 499–510 (2020)
[19] Odierna, B.A.; Silveira, I.F.: Player game data mining for player classification. Proceedings of SBGames. pp. 1–8 (2018)
[20] Jiang, S.; Zhang, L.; Xu, H.; Huang, J.; He, Q.; Zhou, X.; Jiang, J.: GameTrail: Probabilistic lifecycle process model for deep game understanding. Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM). pp. 994–1003 (2024)
[21] Kovačević, M.A.; Pešović, M.D.; Petrović, Z.Z.; Pucanović, Z.S.: Predictive analytics of in-game transactions: Tokenized player history and self-attention techniques. IEEE Access. 12, 1–14 (2024)

[22] Zhao, S.; Xu, Y.; Luo, Z.; Tao, J.; Li, S.; Fan, C.; Pan, G.: Player behavior modeling for enhancing role-playing game engagement. IEEE Transactions on Computational Social Systems. 8(2), 464–474 (2021)
[23] Moon, J.; Choi, Y.; Park, T.; Choi, J.; Hong, J.-H.; Kim, K.-J.: Diversifying dynamic difficulty adjustment agent by integrating player state models into Monte-Carlo tree search. Expert Systems with Applications. 205, 117677 (2022)
[24] Beltagy, I.; Peters, M.E.; Cohan, A.: Longformer: The long-document transformer. ArXiv preprint (2020)
[25] Le, Q.; Mikolov, T.: Distributed representations of sentences and documents. Proceedings of the International Conference on Machine Learning (ICML). pp. 1188–1196 (2014)
[26] Lopes, P.; Yannakakis, G.N.; Liapis, A.: RankTrace: Relative and unbounded affect annotation. Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 158–163 (2017)