Matrix and Tensor Factorization for Profiling Player Behavior

Rafet Sifa

Matrix and Tensor Factorization for Profiling Player Behavior

Rafet Sifa

First version published in January 2019
ISBN: 9781791754570
Imprint: Independently published

For questions, suggestions and typos please write to
mtf4profilingbook@protonmail.com

Updates about the book can be found under
sites.google.com/view/rafetsifa/the-profiling-book

Preface

Matrix and tensor factorization models are scalable and flexible tools for learning efficient representations for a variety of descriptive, predictive and prescriptive analytics tasks. They can be easily deployed without requiring high-performance computing environments and extensive parameter tuning for analyzing large scale and high dimensional behavioral data to come up with interpretable and actionable results. Their range of applications includes recommender systems, behavior prediction, natural language processing, digital forensics, process and budget optimization and behavioral profiling.

In this compact book, we will introduce certain theoretical as well as practical aspects behind a set of matrix and tensor factorization models for behavioral profiling. We will particularly concentrate on the various straightforward-to-implement algorithms used to come up with useful factorizations, ideas to enforce interpretability for human analysts and example behavioral profiling case studies with a behavioral dataset from a digital game. The methods we cover in detail for profiling behavior include unconstrained matrix factorization, Singular Value Decomposition, k-Means clustering, Archetypal Analysis, Simplex Volume Maximization and k-Maxoids Analysis for analyzing bipartite (rectangular) matrices and two- and three-way DEDICOM for respectively decomposing structurally constrained asymmetric similarity matrices and tensors.

Although our main intention is about analyzing player behavior in digital games, the methods we introduced can easily be applied to analyze behavioral telemetry data from similar digital products (such as mobile and web applications) as well. Therefore, this book is suited for developers, analysts, data scientists and engineers, who are interested in tracking, collecting and analyzing behavioral telemetry data to better understand their user-base and automatically learn informative features for a variety of analytics applications.

This work will be based on my recent research on Game Analytics and mainly the second chapter of my PhD thesis [1] with additional content from the first, third, seventh and eighth chapters about the applications. The set of peer-reviewed publications that will mainly accompany this work are from [2–19]. Furthermore, I have published over 40 peer-reviewed papers in the broad area of data science covering purely theoretical as well as application-oriented work in behavioral analytics, psychometrics, text mining and quantum computing. The complete list of my publications can be found on my website `https://sites.google.com/view/rafetsifa`.

Notations

I mostly kept a fixed notation for the equations in this work. Lowercase bold letters symbolize vectors (e.g. $\mathbf{v}$), uppercase bold letters are for matrices (e.g. $\mathbf{V}$) and Chancery font capital letters will stand for tensors (collections of matrices) (e.g. $\mathscr{S}$). Capital letters with calligraphic font will represent mathematical sets (e.g. $\boldsymbol{\mathcal{S}}$) and $|\cdot|$ will represent the set cardinality. Unless stated otherwise, I will be using the column vector notation to specify vectors and the superscript T (e.g. $\mathbf{A}^T$) will be used to represent the transpose of matrices and vectors. I will be mostly using subscripts to index rows, columns and individual elements in vectors, matrices and tensors. For instance, a_i will be the ith element of the vector $\mathbf{a}$, the scalar b_{ij} will be indicating the element in the ith row and the jth column of a matrix $\mathbf{B}$ and c_{ijk} will be pointing to the ith row and the jth column of the kth slice of a tensor $\mathscr{C}$. Additionally $\mathbf{d}_i$ and $\mathbf{d}_{j:}$ will be respectively symbolizing the ith column and the jth row of matrix $\mathbf{D}$. Similarly, $\mathbf{E}_k$ will be referring to the kth slice of a tensor $\mathscr{E}$.

Acknowledgment

I would like to thank my PhD supervisor and mentor Prof. Dr. Christian Bauckhage for his guidance and support for publishing this work as a standalone book. I would like to express my gratitude to Prof. Dr. Anders Drachen for his valuable input and support throughout the last years.

I want to thank Dr. Christian Thurau, Dr. Fabian Hadiji and Prof. Dr. Kristian Kersting for the stimulating research environment we had during my employment at Game Analytics. I would like to thank Julian Runge for his constant support, our joint work and the long discussions that helped me look at the problems from different industry perspectives.

I am also deeply thankful to my current and former colleagues in the Media Engineering department at Fraunhofer IAIS. I want to thank Ulrich Nütten, César Ojeda, Kostadin Cvejoski, Dr. Mirwaes Wahabzada, Eduardo Brito, Ben Wulff, Dr. Anna Ladi, Maren Pielka, Dr. Jannis Schücker and Rajkumar Ramamurthy for the amazing research atmosphere. Warm thanks go to Prof. Dr. Alessandro Canossa, Dr. Sam Devlin, Prof. Dr. Lennart Nacke, Prof. Dr. Tom Stafford, Marc Müller, Prof. Dr. Günter Wallner, Dr. Johanna Pirker, Prof. Dr. Diego Klabjan, Prof. Dr. Pieter Spronck, Yaser Norouzzadeh, Dr. África Periáñez, Paul Bertens, Prof. Dr. Daniel Johnson, Prof. Dr. Fabio Zambetta, Dr. Marco Tamassia, Dr. William Raffe, Prof. Dr. Michael Hitchens, Sridev Srikanth and Raheel Yawar for our valuable discussions and collaborations.

Finally, I would like to thank my family and friends, who have always been with me while preparing this work.

Rafet Sifa
Bonn, Germany
January 2019

Contents

1 Introduction to Player Profiling and Matrix Factorization

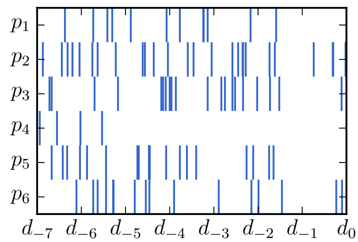
The fundamental objective of the process of profiling user behavior in the context of game analytics is about helping the game development team to better understand player behavior. In this work, we will focus on different matrix and tensor factorization models for factorizing behavioral data matrices to extract efficient representations that reveal interesting insights about the analyzed player population and can be further used in higher-level analytics applications. In order to motivate ourselves to the use of such models in industry settings, in this chapter we will define the notion of behavioral profiling in the context of game analytics, take a brief look at the types of collected behavioral data and the workflow of typical profiling frameworks and finally present an overview of this book.

1.1 Analytics from Telemetry Data

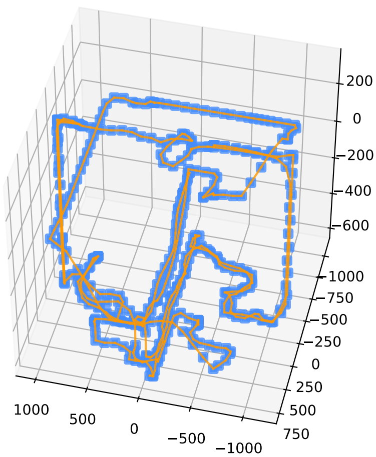
Todays gaming environments are becoming increasingly complex and owing to (arguably) inexpensive storage solutions it is now possible to track every in-game action of any player in real- or quasi-real time. Although the granularity and the type of the collected data is game- and tracking system dependent, there are some types that are universal for most of the settings. These include meta data (e.g. player id, country, registration date, device type etc.), activities on time (e.g. arrivals,

ID	Country	Install
F1SCHER	DE	01/28/2008
300pingpain	US	03/30/2009
kodia034	CN	01/16/2009
MasterBlaster	TR	04/15/2009
br_B0LS0N4R0_br	BR	12/01/2009
CoupDeGrace	BE	02/14/2009
NeverSaw0815	RU	09/11/2009
uragan002	ES	05/05/2008
9Medic	GB	12/08/2009

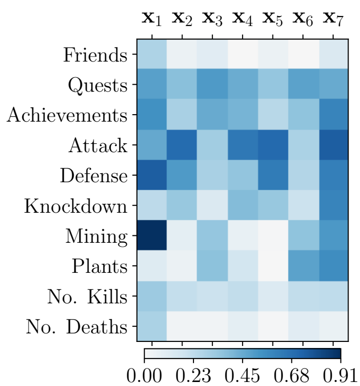
(a) meta information



(b) temporal activities



(c) location



(d) performance snapshots

Figure 1.1: Illustration of different sources of information collected about players of todays games. Common example input types include (a) meta information, (b) activities on time, (c) location in the real and virtual worlds (d) snapshots of predefined performance indicators. Data of these formats can be directly or indirectly fed into matrix and tensor factorization methods to automatically extract features that can be used in a variety of analytics applications. In order to show the behavior of the studied models, in this work, we will be analyzing snapshotted player performance data (similar to (d)). Additionally, we will layout some practical ideas about harnessing different data sources for behavioral profiling at the end of the book.

Bibliography

[1] R. Sifa. *A Machine Learning Perspective on Unraveling, Modeling and Predicting Player Behavior in Games*. PhD thesis, University of Bonn, 2019 (expected).

[2] A. Drachen, R. Sifa, C. Bauckhage, and C. Thurau. Guns, Swords and Data: Clustering of Player Behavior in Computer Games in the Wild,. In *Proc. IEEE CIG*, 2012.

[3] R. Sifa, A. Drachen, C. Bauckhage, C. Thurau, and A. Canossa. Behavior Evolution in Tomb Raider Underworld. In *Proc. IEEE CIG*, 2013.

[4] A. Drachen, C. Thurau, R. Sifa, and C. Bauckhage. A Comparison of Methods for Player Clustering via Behavioral Telemetry. In *Proc. FDG*, 2013.

[5] R. Sifa, C. Bauckhage, and A. Drachen. The Playtime Principle: Large-Scale Cross-Games Interest Modeling. In *Proc. IEEE CIG*, 2014.

[6] R. Sifa, C. Bauckhage, and A. Drachen. Archetypal Game Recommender Systems. In *Proc. KDML-LWA*, 2014.

[7] R. Sifa, C. Ojeda, and C. Bauckhage. User Churn Migration Analysis with DEDICOM. In *Proc. ACM RecSys*, 2015.

[8] C. Bauckhage and R. Sifa. k-Maxoids Clustering. In *Proc. KDML-LWA*, 2015.

[9] C. Bauckhage, A. Drachen, and R. Sifa. Clustering Game Behavior Data. *IEEE Transactions on Computational Intelligence and AI in Games*, 7(3), 2015.

[10] R. Sifa, A. Drachen, and C. Bauckhage. Large-Scale Cross-Game Player Behavior Analysis on Steam. In *Proc. AAAI AIIDE*, 2015.

[11] R. Sifa, S. Srikanth, A. Drachen, C. Ojeda, and C. Bauckhage. Predicting Retention in Sandbox Games with Tensor Factorization-based Representation Learning. In *Proc. IEEE CIG*, 2016.

[12] R. Sifa, C. Ojeda, K. Cvejoski, and C. Bauckhage. Interpretable Matrix Factorization with Stochasticity Constrained Nonnegative DEDICOM. In *Proc. KDML-LWDA*, 2017.

[13] R. Sifa and C. Bauckhage. Online k-Maxoids Clustering. In *Proc. IEEE DSAA*, 2017.

[14] R. Sifa, A. Drachen, and C. Bauckhage. Profiling in Games: Understanding Behavior from Telemetry. In *Social Interaction in Virtual Worlds*. Cambridge University Press, 2018.

[15] R. Sifa. An Overview of Frank-Wolfe Optimization for Stochasticity Constrained Interpretable Matrix and Tensor Factorization. In *Proc. ICANN*. Springer, 2018.

[16] C. Bauckhage, K. Kersting, R. Sifa, C. Thurau, A. Drachen, and A. Canossa. How Players Lose Interest in Playing a Game: An Empirical Study Based on Distributions of Total Playing Times. In *Proc. IEEE CIG*, 2012.

[17] R. Sifa and C. Bauckhage. Archetypical Motion: Supervised Game Behavior Learning with Archetypal Analysis. In *Proc. IEEE CIG*, 2013.

[18] C. Bauckhage, R. Sifa, A. Drachen, C. Thurau, and F. Hadiji. Beyond Heatmaps: Spatio-Temporal Clustering Using Behavior-Based Partitioning of Game Levels. In *Proc. IEEE CIG*, 2014.

[19] A. Rattinger, G. Wallner, A. Drachen, J. Pirker, and R. Sifa. Integrating and Inspecting Combined Behavioral Profiling and Social Network Models in Destiny. In *Proc. IFIP ICEC*. Springer, 2016.

[20] El-Nasr, M.S. and Drachen, A. and Canossa, A. *Game Analytics: Maximizing the Value of Player Data*. Springer, 2013.

[21] J. Runge, P. Gao, F. Garcin, and B. Faltings. Churn Prediction for High-Value Players in Casual Social Games. In *Proc. IEEE CIG*, 2014.

[22] Newzoo. The Mobile Games Market: Casual Games Sector Report. Technical report, 2016.

[23] Y. Koren. The Bellkor Solution to the Netflix Grand Prize. *Netflix Prize Documentation*, 81, 2009.

[24] G. Takács, I. Pilászy, B. Németh, and D. Tikk. Scalable Collaborative Filtering Approaches for Large Recommender Systems. *Journal of Machine Learning Research*, 10, 2009.

[25] C. Bauckhage. k-Means Clustering via the Frank-Wolfe Algorithm. In *Proc. KDML-LWDA*, 2016.

[26] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

[27] M. Ahmad, B. Keegan, J. Srivastava, D. Williams, and N. Contractor. Mining for Gold Farmers: Automatic Detection of Deviant Players in MMOGs. In *Proc. IEEE CSE*, 2009.

[28] A. Cutler and L. Breiman. Archetypal Analysis. *Technometrics*, 36(4), 1994.

[29] C. Thurau, K. Kersting, and C. Bauckhage. Yes We Can: Simplex Volume Maximization for Descriptive Web-Scale Matrix Factorization. In *Proc. ACM CIKM*, 2010.

[30] R. Harshman. Models for Analysis of Asymmetrical Relationships among N Objects or Stimuli. In *Proc. Joint Meeting of the Psychometric Society and the Society for Mathematical Psychology*, 1978.

[31] H. Kiers. DESICOM: Decomposition of Asymmetric Relationships Data into Simple Components. *Behaviormetrika*, 24(2), 1997.

[32] L. Trefethen and D. Bau. *Numerical Linear Algebra*, volume 50. SIAM, 1997.

[33] G. Golub and C. Van Loan. *Matrix Computations*, volume 3. JHU Press, 2012.

[34] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Application of Dimensionality Reduction in Recommender System- A Case Study. Technical Report TR 00-043, University of Minnesota Department of Computer Science, 2000.

[35] Y. Koren, R. Bell, and C. Volinsky. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer*, 42(8), 2009.

[36] E. Anderson, Z. Bai, J. Dongarra, A. Greenbaum, A. McKenney, J. Du Croz, S. Hammarling, J. Demmel, C. Bischof, and D. Sorensen. LAPACK: A Portable Linear Algebra Library for High-performance Computers. In *Proc. ACM/IEEE Conf. on Supercomputing*, 1990.

[37] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Philip, et al. Top 10 Algorithms in Data Mining. *Knowledge and Information Systems*, 14(1), 2008.

[38] A. Coates and A. Ng. Learning Feature Representations with k-Means. In *Neural networks: Tricks of the Trade*. Springer, 2012.

[39] C. Bauckhage, E. Brito, K. Cvejoski, C. Ojeda, R. Sifa, and S. Wrobel. Adiabatic Quantum Computing for Binary Clustering. *arXiv preprint arXiv:1706.05528*, 2017.

[40] S.P. Lloyd. Least Squares Quantization in PCM. *IEEE Trans. Information Theory*, 28(2), 1982.

[41] C. Bauckhage and C. Thurau. Making Archetypal Analysis Practical. In *Proc. DAGM Symposium on Pattern Recognition*, 2009.

[42] M. Morup and L.K. Hansen. Archetypal Analysis for Machine Learning and Data Mining. *Neurocomputing*, 80(Mar):54–63, 2012.

[43] M. Frank and P. Wolfe. An Algorithm for Quadratic Programming. *Naval Research Logistics Quarterly*, 3(1-2), 1956.

[44] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient Projections onto the l1-Ball for Learning in High Dimensions. In *Proc. ACM ICML*, 2008.

[45] M. Jaggi. Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. In *Proc. ACM ICML*, 2013.

[46] R. Sifa, D. Paurat, D. Trabold, and C. Bauckhage. Simple Recurrent Neural Networks for Support Vector Machine Training. In *Proc. ICANN*. Springer, 2018.

[47] C. Thurau, K. Kersting, M. Wahabzada, and C. Bauckhage. Descriptive Matrix Factorization for Sustainability: Adopting the Principle of Opposites. *Data Mining and Knowledge Discovery*, 24(2), 2012.

[48] C. Faloutsos and K.I. Lin. FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. In *Proc. SIGMOD*, 1995.

[49] J.B. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In *Proc. Berkeley Symp. on Mathematical Statistics and Probability*, 1967.

[50] T.M. Martinetz, S.G. Berkovich, and K. Schulten. Neural Gas Network for Vector Quantization and its Application to Time Series Prediction. *IEEE Trans. Neural Networks*, 4(4), 1993.

[51] T. Kohonen. The Self-Organizing Map. *Neurocomputing*, 21 (1–3), 1998.

[52] R. A. Fisher and L. H. C. Tippett. Limiting Forms of the Frequency Distribution of the Largest of Smallest Member of a Sample. *Proc. Cambridge Philosophical Society*, 24(2), 1928.

[53] B. Gnedenko. Sur la Distribution Limite du Terme Maximum d'une Série Aléatoire. *Annals of Mathematics*, 44(3), 1943.

[54] S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer London, 2001.

[55] C. Lai, M. Xie, and D. Murthy. A Modified Weibull Distribution. *IEEE Transactions on Reliability*, 52(1), 2003.

[56] W. Weibull. A Statistical Distribution Function of Wide Applicability. *Journal of Applied Mechanics*, 103(730), 1951.

[57] H. Rinne. *The Weibull Distribution: A Handbook*. Chapman and Hall/CRC, 2008.

[58] F. R. Nash. *Reliability Assessments: Concepts, Models, and Case Studies*. CRC Press, 2016.

[59] C. Bauckhage and K. Kersting. Collective Attention on the Web. *Foundations and Trends in Web Science*, 5(1-2), 2016.

[60] C. Chambers, W. Feng, S. Sahu, and D. Saha. Measurement-Based Characterization of a Collection of On-line Games. In *Proc. ACM SIGCOMM Conf. on Internet Measurement*, 2005.

[61] B.W. Bader, R. Harshman, and T.G. Kolda. Temporal Analysis of Semantic Graphs Using ASALSAN. In *Proc. IEEE ICDM*, 2007.

[62] M. Nickel, V. Tresp, and H. Kriegel. A Three-Way Model for Collective Learning on Multi-Relational Data. In *Proc. ACM ICML*, 2011.

[63] C. Lawson and R. Hanson. *Solving Least Squares Problems*. SIAM, 1974.

[64] R. Bro and S. De Jong. A Fast Non-negativity Constrained Least Squares Algorithm. *Journal of Chemometrics*, 11(5), 1997.

[65] P. Kantor, L. Rokach, F. Ricci, and B. Shapira. *Recommender Systems Handbook*. Springer, 2011.

[66] P. Resnick and H. Varian. Recommender Systems. *Communications of the ACM*, 40(3), 1997.

[67] G. Linden, B. Smith, and J. York. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, 7(1), 2003.

[68] P. Cremonesi, Y. Koren, and R. Turrin. Performance of Recommender Algorithms on Top-N Recommendation Tasks. In *Proc. ACM Recsys*, 2010.

[69] R. Sifa, R. Yawar, R. Ramamurthy, and C. Bauckhage. Matrix and Tensor Factorization based Game Content Recommender Systems: A Bottom-Up Architecture and a Comparative Online Evaluation. In *Proc. AAAI AIIDE*, 2018.

[70] B. Weber. Building A Recommendation System for EverQuest Landmark's Marketplace. In GDC, 2015.

[71] P. Bertens, A. Guitart, P. P. Chen, and Á. Periáñez. A Machine-Learning Item Recommendation System for Video Games. *arXiv preprint arXiv:1806.04900*, 2018.

[72] L. Hanke and L. Chaimowicz. A Recommender System for Hero Line-ups in MOBA Games. In *Proc. AAAI AIIDE*, 2017.

[73] R. Sifa, E. Pawlakos, K. Zhai, S. Haran, R. Jha, D. Klabjan, and A. Drachen. Controlling the Crucible: A Novel PvP Recommender Systems Framework for Destiny. In *Proc. ACM ACSW IE*, 2018.

[74] A. Drachen, R. Sifa, and C. Thurau. The Name in the Game: Patterns in Character Names and Gamer Tags. *Entertainment Computing*, 5(1), 2014.

[75] A. Kokkinakis, J. Lin, D. Pavlas, and A. Wade. What's in a Name? Ages and Names Predict the Valence of Social Interactions in a Massive Online Game. *Computers in Human Behavior*, 55, 2016.

[76] S. C. Herring, D. O. Kutz, J. Paolillo, and A. Zelenkauskaite. Fast Talking, Fast Shooting: Text Chat in an Online First-Person Game. In *Proc. IEEE HICSS*, 2009.

[77] E. Brito, R. Sifa, K. Cvejoski, C. Ojeda, and C. Bauckhage. Towards German Word Embeddings: A Use Case with Predictive Sentiment Analysis. In *Data Science–Analytics and Applications*. Springer, 2017.

[78] J. O. Ryan, E. Kaltman, M. Mateas, and N. Wardrip-Fruin. What We Talk About When We Talk About Games: Bottom-up Game Studies Using Natural Language Processing. In *Proc. FDG*, 2015.

[79] C. Drescher, G. Wallner, S. Kriglstein, R. Sifa, A. Drachen, and M. Pohl. What Moves Players?: Visual Data Exploration of Twitter and Gameplay Data. In *Proc. Conference on Human Factors in Computing Systems*. ACM, 2018.

[80] T. K. Landauer, P. W. Foltz, and D. Laham. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2-3), 1998.

[81] J. Kawale, A. Pal, and J. Srivastava. Churn Prediction in MMORPGs: A Social Influence Based Approach. In *Proc. IEEE CSE*, 2009.

[82] J. Pirker, A. Rattinger, A. Drachen, and R. Sifa. Analyzing Player Networks in Destiny. *Entertainment Computing*, 25, 2018.

Index