Information Systems Frontiers (2018) 20:1153–1156
https://doi.org/10.1007/s10796-018-9884-1

Guest Editorial: Behavioral-Data Mining in Information Systems and the Big Data Era

Ludovico Boratto[1] · Salvatore Carta[2] · Andreas Kaltenbrunner[3,4] · Matteo Manca[5]

Published online: 25 October 2018

1 Introduction

An information system collects and processes data with the aim to extract information and to support decision-making tasks. Since the advent of the so-called *Social Web*, users are encouraged to create content and upload it on the Web, so huge amounts of data are continuously generated. This data represents a great opportunity for researchers, companies, and decision makers to infer non-trivial patterns and generate new knowledge; on the other side, a lot of challenges arise from such amounts of data. In order to handle these new challenges and accomplish their objectives, information systems need efficient and effective ways to process this data. On the one hand, the algorithms that process these large amounts of data should have low computational costs, in order to keep up with the rapid evolution of the Web and guarantee efficiency, while on the other hand they should be able to filter out the less useful chunks of data and process only those that lead to an effective decision making.

Behavioral-data mining is the process of extracting information by analyzing the huge amounts of data that describe the behavior of the users in a system. This particular kind of mining has proven to be useful in various information systems areas (Beutel et al. 2015), such as the detection of tag clusters (Boratto et al. 2009), the creation of web personalization services (Mobasher et al. 2000), the improvement of web search ranking (Agichtein et al. 2006), and the generation of friend recommendations in social media systems (Manca et al. 2018). It is also the foundation of many computational social science studies (Lazer et al. 2009).

In this special issue, we explore a new frontier in Information Systems, which aims at producing behavioral-data mining approaches able to deal with the big data problem. The rest of this article is structured as follows: Section 2 focuses on the challenges that mining behavioral data in big data scenarios poses; in Section 3 we introduce some recent advances in this area; Section 4 contains concluding remarks.

2 Research Challenges

Mining behavioral data in information systems, when working in big data scenarios, poses several challenges. In this section, we will discuss some of the most important ones.

When monitoring users' behavior, we mostly rely on *implicit feedback* provided by the users (e.g., the browsing history, or the items the users click on) (Oard and Kim 1998). While this form of feedback allows us to collect much more information about the users with respect to explicit feedback provided by ratings or reviews, its drawback is the lack of information about what the users do not like. Indeed, missing feedback might mean that the user does not like an item or that she might have not encountered it. Thinking about domains such as large e-commerce websites or social media platforms, in which users interact only with a small subset of items, this leads to

✉ Ludovico Boratto
ludovico.boratto@acm.org

Salvatore Carta
salvatore@unica.it

Andreas Kaltenbrunner
kaltenbrunner@gmail.com

Matteo Manca
matteomanca@gmail.com

[1] Eurecat, Centre Tecnològic de Catalunya, Barcelona, Spain
[2] University of Cagliari, Cagliari, Italy
[3] NTENT, Barcelona, Spain
[4] Universitat Pompeu Fabra, Barcelona, Spain
[5] Zurich, Barcelona, Spain

a lot of uncertainty on the preferences for the majority of the items. In the mining process, this might affect the extraction of actionable knowledge about the users.

Both when working with implicit and explicit feedback, data is very sparse. Indeed, as previously mentioned, users implicitly interact with a small subset of items; moreover, they are usually reluctant to provide explicit ratings to evaluate the items, which is considered as a tiring process (Oard and Kim 1998). As Fan et al. highlight (Fan et al. 2014), high dimensionality leads to noise accumulation, spurious correlations, and incidental homogeneity. Moreover, when combining high-dimensional data to high sparsity, issues such as heavy computational cost and algorithmic instability arise. Hence, processing behavioral data still represents a challenge.

In the growing field of Computational Social Science, solving these problems has a direct impact in how science is done, how large scale sociological experiments have to be designed, or how given data sets of "natural experiments" (Dunning 2012) can be exploited scientifically to allow to verify or refute hypotheses. Challenges in this context are control for potential confounders in observational data, for example through randomization of the collected data. This allows to compare the observed with the expected outcome and assess the degree to which this outcome is indeed relevant. An example for such a study is given in Laniado et al. (2018). In case one is in control of the data collection process prior to the observed events, the additional challenge of how to correctly design the data collection process has to be solved, as is shown in the case for urban mobility studies in Manca et al. (2017).

Last, but not least, users' privacy is a very timely issue, considering the new *General Data Protection Regulation (GDPR)*, which is enforceable throughout Europe since May 25, 2018. Note that Facebook decided to extend the regulation worldwide,[1] so the services that are built on top of it have to be compliant with the regulation, independently from the country. In addition to how data is collected and stored, according to Article 13, Paragraph 2 (f) of GDPR, users are entitled to have an explanation about how decisions are taken by an algorithm. Hence, collecting and storing behavioral data after the regulation might be challenging, and the mining algorithms should be able to provide explanations to users, which might not always be possible (e.g., when employing deep learning).

3 In this Special Issue

The six accepted articles in this special issue cover many of the aforementioned themes, with innovative techniques for mining user behavior in information systems in big data environments. The research contributions advance the state of the art, considering different aspects and scenarios, ranging from content retrieval and classification, to the characterization of user satisfaction in social networks and recommender systems, by considering different aspects, such as user personality, geographic distance of the users, and the content of the items.

In their article "TV-Program Retrieval and Classification: A Comparison of Approaches based on Machine Learning", Narducci et al. (2018) analyze user behavior in order to generate personalized Electronic Program Guides (EPGs). More specifically, they focus on the retrieval of possibly interesting programs for the users, by first classifying them according to their textual description, and then retrieving those that best match a specific program type. Experiments performed on a dataset provided by Philips Research, related to 133,579 TV shows broadcasted by 47 channels in German language, show that Logistic Regression is the best algorithm both in the classification and retrieval tasks.

Nguyen et al. considered the role of personality in recommender systems, in the paper "User Personality and User Satisfaction with Recommender Systems" (Nguyen et al. 2018). This study considers 1800 users, to analyze if rating-based recommender systems were able to deliver preferred levels of diversity, popularity, and serendipity to them. Results show that these systems fail to do so. The authors also assessed users' personality traits using the Ten-item Personality Inventory (TIPI), which suggests that users with different personalities have different preferences for these three recommendation properties. Given these results, the authors suggest that, in the future, recommender systems should consider users' personality traits.

Golbeck et al., in their article "Scaling Up Integrated Structural and Content-Based Network Analysis" (Golbeck et al. 2018), face the issue of identifying clusters and classifying network nodes, as the network grows bigger and manual classification is no longer possible. The authors show how topic modeling can be employed to produce easy-to-understand keywords that represent important clusters in a network. Those keywords reflect the insights achieved by human analysts doing a manual content-based analysis of the network features.

The paper "The Impact of Geographic Distance on Online Social Interactions", by Laniado et al. (2018), aims to explain the effect that geographic distance has on online social interactions and, simultaneously, tries to understand the interplay between the social characteristics of friendship ties and their spatial properties. The findings support the idea that spatial distance constraints whom users interact with, but not the intensity of their social interactions. Furthermore, friendship ties belonging to denser connected groups tend to arise at shorter spatial distances than social

[1] https://newsroom.fb.com/news/2018/04/new-privacy-protections/

ties established between members belonging to different groups. Finally, the authors show that these findings mostly do not depend on the age of the users, although younger users seem to be slightly more constrained to shorter geographic distances.

In their article "Inducing Personalities and Values from Language Use in Social Network Communities", Kumar et al. (2018) analyze the communities in social media networks as composition of induced psycholinguistic and sociolinguistic variables (Personalities, Values, and Ethics) across individuals. The study was performed on six datasets annotated with Values and Ethics of the users. The authors created models to determine the Personality and Values of individuals, by analyzing their language usage and social media behavior. Then, they connect the characteristics of individuals within an online community, and they create a map of values and ethics for India.

The final paper of this special issue, titled "Personality, User Preferences and Behavior in Recommender Systems", by Karumur et al. (2018), identified Big-5 personality types of 1840 users of the MovieLens recommender system. The aim was to examine factors of user retention and engagement, content preferences, and rating patterns, to identify recommender-system related behaviors and preferences that correlate with user personality. Results show that personality traits correlate significantly with behaviors and preferences such as newcomer retention, intensity of engagement, activity types, item categories, consumption versus contribution, and rating patterns.

4 Conclusions

Mining user behavior in information systems is a topic of central interest to gather actionable knowledge about the users and provide services to them. Being able to do so in scenarios characterized by the big data represents a new frontier in this area. The papers included in this special issue cover several topics and present some of the key directions in this vibrant and rapidly expanding area of research and development. We hope the set of selected papers provides the community with a better understanding of the current directions, and that they inspire readers with possible areas to focus on in their future research.

Acknowledgments We thank all the authors for considering this special issue as an outlet to publish their research results in the area of behavioral-data mining. We also would like to thank the referees who provided very useful and thoughtful feedback to the authors. Finally, we express our gratitude to the Editors-in-Chief, Professor Rao and Professor Ramesh, for their kind support, advice, and encouragements throughout the preparation of this special issue.

References

Agichtein, E., Brill, E., Dumais, S.T. (2006). Improving web search ranking by incorporating user behavior information. In E.N. Efthimiadis, S.T. Dumais, D. Hawking, K. Jarvelin (Eds.), *SIGIR 2006: proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, Seattle, Washington, August 6–11, 2006* (pp. 19–26). ACM. https://doi.org/10.1145/1148170.1148177.

Beutel, A., Akoglu, L., Faloutsos, C. (2015). Graph-based user behavior modeling: from prediction to fraud detection. In L. Cao, C. Zhang, T. Joachims, G.I. Webb, D.D. Margineantu, G. Williams (Eds.), *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, Sydney, NSW, Australia, August 10–13, 2015* (pp. 2309–2310). ACM. https://doi.org/10.1145/2783258.2789985.

Boratto, L., Carta, S., Vargiu, E. (2009). RATC: a robust automated tag clustering technique. In T.D. Noia, & F. Buccafurri (Eds.), *E-Commerce and web technologies, 10th international conference, EC-Web 2009, Linz, Austria, September 1–4, 2009. Proceedings, Lecture Notes in Computer Science* (Vol. 5692, pp. 324–335). Springer, https://doi.org/10.1007/978-3-642-03964-5_30.

Dunning, T. (2012). *Natural experiments in the social sciences: a design-based approach*. Cambridge: Cambridge University Press.

Fan, J., Han, F., Liu, H. (2014). Challenges of big data analysis. *National Science Review*, *1*(2), 293–314. https://doi.org/10.1093/nsr/nwt032.

Golbeck, J., Gerhard, J., O'Colman, F., O'Colman, R. (2018). Scaling up integrated structural and content-based network analysis. *Information Systems Frontiers, 20*(6). https://doi.org/10.1007/s10796-017-9783-x.

Karumur, R.P., Nguyen, T.T., Konstan, J.A. (2018). Personality, user preferences and behavior in recommender systems. *Information Systems Frontiers, 20*(6). https://doi.org/10.1007/s10796-017-9800-0.

Kumar, U., Reganti, A.N., Maheshwari, T., Chakroborty, T., Gambäck, B., Das, A. (2018). Inducing personalities and values from language use in social network communities. *Information Systems Frontiers, 20*(6). https://doi.org/10.1007/s10796-017-9793-8.

Laniado, D., Volkovich, Y., Scellato, S., Mascolo, C., Kaltenbrunner, A. (2018). The impact of geographic distance on online social interactions. *Information Systems Frontiers, 20*(6). https://doi.org/10.1007/s10796-017-9784-9.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., Van Alstyne, M. (2009). Computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742.

Manca, M., Boratto, L., Roman, V.M., i Gallissà, O.M., Kaltenbrunner, A. (2017). Using social media to characterize urban mobility patterns: state-of-the-art survey and case-study. *Online Social Networks and Media*, *1*, 56–69. https://doi.org/10.1016/j.osnem.2017.04.002.

Manca, M., Boratto, L., Carta, S. (2018). Behavioral data mining to produce novel and serendipitous friend recommendations in a social bookmarking system. *Information Systems Frontiers*, *20*(4), 825–839. https://doi.org/10.1007/s10796-015-9600-3.

Mobasher, B., Cooley, R., Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM*, *43*(8), 142–151. https://doi.org/10.1145/345124.345169.

Narducci, F., Musto, C., de Gemmis, M., Lops, P., Semeraro, G. (2018). Tv-program retrieval and classification: a comparison of approaches based on machine learning. *Information Systems Frontiers, 20*(6). https://doi.org/10.1007/s10796-017-9780-0.

Nguyen, T.T., Maxwell Harper, F., Terveen, L., Konstan, J.A. (2018). User personality and user satisfaction with recommender systems. *Information Systems Frontiers, 20*(6). https://doi.org/10.1007/s10796-017-9782-y.

Oard, D., & Kim, J. (1998). Implicit feedback for recommender systems. In *Proceedings of the AAAI workshop on recommender systems* (pp. 81–83).

Ludovico Boratto is a research scientist in the Data Science and Big Data Analytics research group at Eurecat, in Barcelona (Spain). His research interests focus on Data Mining and Machine Learning approaches, mostly applied to recommender systems and social media analysis. He is editor of the book "Group Recommender Systems: An Introduction", published by Springer. In 2012, he got a Ph.D. at the University of Cagliari (Italy), where he was research assistant until May 2016. In 2010 and 2014 he spent 10 months at Yahoo! Research in Barcelona as a visiting researcher. He is member of the ACM and of the IEEE.

Salvatore Carta is Associate Professor in Computer Science at the University of Cagliari (Italy). His current research interests include behavioral pattern identification and recommendation, AI algorithms for credit scoring, fraud detection, intrusion detection, financial forecasting and robo-trading, and e-coaching platforms for healthy lifestyles. He is co-founder of the"Trustworthy Computational Societies Research Group" and of the "Artificial Intelligence and Big Data Group" at the Mathematics and Computer Science Department of the University of Cagliari. He holds a PhD in Electronics and Computer Science from the University of Cagliari. He is member of the ACM and of the IEEE.

Andreas Kaltenbrunner is Director of Data Analytics at NTENT, where he leads a team focusing on user behavior analysis and improvements for ranking in mobile search with the aim of increasing users satisfaction and retention. Andreas is also teaching a master course on Data Driven Social Analytics and is involved in research activities centered on computational social science, social media and social network analysis, areas in which he has co-authored more than 60 publications. He obtained his PhD in Computer Science and Digital Communication in 2008 from the Universitat Pompeu Fabra with a thesis about stochastic effects in human and neural communication patterns. Afterwards, he joined as researcher the technology center Barcelona Media, where he led from 2013 onwards the Social Media Research Line. Between June 2015 and August 2017 he lead the Digital Humanities Research Unit at the technology center Eurecat, before joining NTENT in September 2017.

Matteo Manca is data scientist at Zurich, where his main activities are related to the study, implementation and validation of predictive and statistical models and to the analysis of data aimed at quantifying risk and supporting business decision making. He previously was data scientist in the Data Science and Big Data Analytics research group at Eurecat, where he mainly focused on the analysis of digital trace data and in the application of data mining and computation methods in order to study social phenomena. In 2014 he obtained his PhD in computer science from the University of Cagliari with a thesis focused on the study and implementation of social recommendation approaches for the social media domain.