Human-recommender Systems: From Benchmark Data to Benchmark Cognitive Models *

Patrick Shafto
Rutgers University – Newark
Newark, NJ 07102
patrick.shafto@gmail.com

Olfa Nasraoui
University of Louisville
Louisville, KY
olfa.nasraoui@louisville.edu

ABSTRACT

We bring to the fore of the recommender system research community, an inconvenient truth about the current state of understanding how recommender system algorithms and humans influence one another, both computationally and cognitively. Unlike the great variety of supervised machine learning algorithms which traditionally rely on expert input labels and are typically used for decision making by an expert, recommender systems specifically rely on data input from non-expert or casual users and are meant to be used directly by these same non-expert users on an every day basis. Furthermore, the advances in online machine learning, data generation, and predictive model learning have become increasingly interdependent, such that each one feeds on the other in an iterative cycle. Research in psychology suggests that people's choices are (1) contextually dependent, and (2) dependent on interaction history. Thus, while standard methods of training and assessing performance of recommender systems rely on benchmark datasets, we suggest that a critical step in the evolution of recommender systems is the development of benchmark *models* of human behavior that capture contextual and dynamic aspects of human behavior. It is important to emphasize that even extensive real life user-tests may not be sufficient to make up for this gap in benchmarking validity because user tests are typically done with either a focus on user satisfaction or engagement (clicks, sales, likes, etc) with whatever the recommender algorithm suggests to the user, and thus ignore the human cognitive aspect. We conclude by highlighting the interdisciplinary implications of this endeavor.

1. INTRODUCTION

In its pioneering years, machine learning drew significantly from cognitive psychology and human learning. Later, with the proliferation of online services, social media websites, and more generally, the wide democratization of "consumption of algorithms' output" by general users, machine learning algorithms started interacting with users at unprecedented rates. While in the early years, most (supervised) machine learning algorithms relied on reliable expert labels to build predictions [23, 24, 25, 8], more recently the gates of data generation have been opened wide to the general population with everyday users' interactions–labeling, rating, annotating, etc–being treated as training data for subsequent interactions and users [6, 19, 11].

A critical development in the field of machine learning, and recommender systems specifically, has been the curation of benchmark datasets. Prominent recent examples include the UCI respository [18], MNIST [17], as well as many others. Benchmark datasets facilitate advances in the field by promoting consensus regarding the scope of the problems to be solved and by offering standards by which all models can be assessed. Indeed, the importance of these datasets to the field is reflected in the high number of citations to these resources—more than 1700 for UCI and more than 500 for MNIST, which almost certainly under-represent the frequency of their use and their importance to the field.

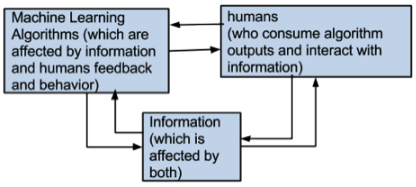


Figure 1: Illustration of the interaction between humans and algorithms through which data are generated. The arrows indicate directions of influence. Traditional recommender systems research focuses on the links between algorithms and data through curated benchmark data sets. Increasingly, algorithms are deployed in the "wild" and iteratively refined based on uncurated data, forming human-recommender systems. This raise questions about the non-independence of data and context and the non-stationarity of behavior over time.

Benchmark datasets typically provide data in batch format. This reflects the historical approach to machine learning and recommendation in which an algorithm is trained on a batch of data and then deployed. It also reflects assumptions about the nature of the problem to be solved; the tar-

*This research was supported in part by NSF grant NSF-1549981 to O.N. and P.S.

RecSys '16, September 15 - 19, 2016, Boston , MA, USA

ISBN 978-1-4503-4035-9/16/09…$15.00

DOI: http://dx.doi.org/10.1145/2959100.2959188

get distribution is stationary and that data are IID samples. These are fundamental assumptions of most standard learning algorithms, and thus sensible simplifying assumptions for starting out. The success of these standard algorithms and approaches has, however, led to new, more ambitious applications including online information filtering and item recommendations, retraining of algorithms based on users' input, and iteration of these interactions. Implicitly, iterative retraining with algorithms that assume stationarity and IID data leads to the potential for introduction of biases, which may be magnified through the interactions over time.

In light of these developments, we argue that it is necessary that the field reconsider how it trains and assesses algorithms. We propose the development of *benchmark models* of human behavior. Analogous to benchmark data sets, benchmark models capture the generative process that underlies typical benchmark data, including time-varying aspects. Benchmark models will accelerate progress by encouraging consensus regarding the kinds of dynamics and non-independence that are important for recommendation, and minimize risk by formalizing the aspects human behavior that potentially introduce bias and allowing researchers to investigate implications for their own algorithms. We emphasize that while some previous research has considered the dynamics of data into model evaluation [26], these have looked primarily at Benchmark data as a dynamic data stream, and not at the interactions between algorithms and humans.

We begin by stating minimal desiderata for a benchmark models vis-a-vis benchmark data. We then review key elements of human behavior to be integrated into such a model: contextual dependence of choice behavior, non-stationarities of preference due to learning about the domain of recommendation, and non-stationarities of preference due to learning about the recommender system itself. We conclude by summarizing our interactive approach and implications for the development of true human-recommender systems.

2. DESIDERATA FOR BENCHMARK MODELS

There are two minimal desiderata for benchmark models of human behavior.

- Marginalizing over time, the data produced by the benchmark model should correspond to traditionally curated benchmark data used in research on recommender systems and machine learning.
- The temporal dynamics of the model should be time varying such that it captures contextual-dependencies and non-stationarities observed in behavioral research and models of human behavior.

The first aspect is, we expect, relatively uncontroversial. Indeed, the simplest version of this, which does not include temporal dynamics, is precisely the generative modeling approaches that have been popular in probabilistic modeling.

The second aspect is, we expect, potentially more controversial along two dimensions: is this necessary and can this be done? We believe there is a very compelling reason for the necessity of models that capture dynamics of human behavior. If behavior is time varying, no batch method can capture the statistical structure of this behavior. This necessarily implies a deviation between the algorithms, which assume and are trained on data that does not possess interesting temporal structure, and people whose behavior is time varying. The iterated nature of interactions means that, under some circumstances, this gap may grow over time, suggesting that it is possible to observe a divergence between algorithms and behavior that grows over time. Benchmark models that capture the dynamics of of human behavior would allow one to test algorithms for satisfactory behavior both on average and across time. In the next section, we address the second question, whether such benchmark models are feasible by considering the elements of a model of human behavior.

3. ELEMENTS OF A MODEL OF HUMAN BEHAVIOR

The main goal of the model of human behavior is to abstract away from benchmark datasets to generative models for how people produce the data. This more abstract framework can allow the sorts of non-stationarity and non-independence that have been identified in research in cognitive science.

Research into human choice behavior in cognitive science dates back at least to Luce [20, 22, 35, 31]. Luce formalized human choice through the eponymous Luce choice rule, in which the probability of choosing an option from a set of possibilities is proportional to its utility. Luce's choice rule implicitly assumes a form of independence, independence from irrelevant alternatives. Since his proposal, researchers have identified numerous examples of how people's choices do in fact depend on what one would reasonably consider irrelevant alternatives. For example, several of the most famous examples stem from the work of Kahneman and Tversky [34, 29]. Consider the attraction effect, in which, people choose between, a nice pen or 6 dollars. At baseline, people favored the money over the pen; however, adding a second, much less attractive pen led more people to choose the nice pen. Adding an irrelevant pen increased participants preference for the nicer pen as compared to the money. This effect is a demonstration of the non-independence of choice in that the order of people's preferences is not stable with respect to the addition of other options (as is required by independence). Moreover, inspired by this work, there are a large number of computational models of choice that have been proposed to account for these non-independence effects, which would provide candidate starting points for formalizing a benchmark model of human behavior [22, 35, 33].

Increasingly, research in cognitive science has also investigated avenues by which people's choice behavior would be non-stationary over time. Two main mechanisms of non-stationarity are the active learning about the information they are searching for and learning about the system itself. For active learning, non-stationarities would arise due to people's beliefs changing over time. Thus, a choice at time $t+1$ may differ from that at $t-1$ only because people learned more about the options at time t. Active learning has a large body of research in cognitive science dating back at least to Bruner [9]. Recent research has focused on developing computational accounts of active learning [10, 3, 1, 2, 21], with recent work focusing on information gain (aka, KL divergence) as a model of human behavior in simple settings.

A second potential source of non-stationarity is people learning about the system with which they are interacting.

These non-stationarities arise in that people may develop heuristics to influence the system toward their goals. Here non-stationarities arise because people are learning about the behavior of the system itself, in essence reverse engineering their input to influence the machine toward their desired goals. This notion has an analog in recent cognitive science investigating learning about, and from, other people [28]. These approaches extend models of choice to incorporate simple models of other people's actions in terms of plans [4] and through more abstract modeling of the individual's goals and desires [13, 5, 28, 27, 12].

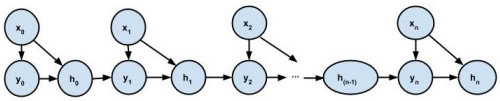
Although research continues, there are reasons to believe that development of a model of human behavior is reasonable. The first is that the basic empirical phenomena in choice theory are agreed upon. Second, the approaches based on learning are being developed within the broad choice theoretic framework [5, 28], suggesting a continuity in modeling.

4. ITERATED LEARNING

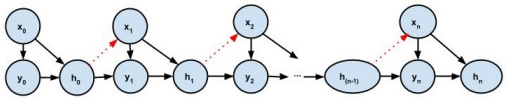
We argue for a framework for investigating the implications of interactions between human and algorithms that draws on diverse literature to provide algorithmic, mathematical, computational, and behavioral tools for investigating human-algorithm interaction. Such an approach would draw on foundational algorithms for selecting and filtering of data from computer science, while also adapting mathematical methods from the study of cultural evolution [14, 16, 7] to formalize the implications of iterative interactions. The basic feature of the algorithm application "in the wild" is iterative interaction with people. In this sense, the evolution of algorithms is a special case of cultural evolution of the sort observed in human language [15] and human knowledge more broadly [32, 7]. Researchers in the behavioral sciences have developed formal and empirical frameworks for analyzing and investigating the asymptotic effects of iterative interactions. Formally, iterative interactions with transmission between adjacent iterations forms a Markov chain [14]. Markov chains provide a well-elaborated framework for analyzing the effects of local decisions on long-run behavior. For the current proposal, the key question regards the conditions under which algorithms' behavior—in terms of the choice of data to present to people and the updated behavior in response to people's observed actions—will converge to more or less effective performance in the long run. Can we understand and compensate for these biases to ensure good performance?

Consider simple supervised machine learning where the goal is to learn to predict a discrete class label. Research on simple classifiers has historically paved the road for most formal analysis in machine learning and data mining, and had a significant impact on both information retrieval [30] as well as recommender systems [6]. It also provides a point of contact with the psychology of category learning.

We can extend iterated learning to provide a framework to analyze the evolution of the learned hypotheses **with interactions between the user and the algorithm**. We will account for the dependency between the current hypothesis and the next input because the model or hypothesis learned by the algorithm (learner) is used as a filter or gateway to the types of data that will later **be seen by** the user. This constitutes dependence between the current hypotheses h and the next inputs $\mathbf{x}$ (see Fig 2). It is interesting to contrast the evolution of iterated learning without and with this dependency. Without the dependency, the algorithm at step $n+1$ sees input $\mathbf{x}_{n+1}$ which is generated from a distribution $p(\mathbf{x})$ that is independent of all other variables. Represent this independence with new notation $q(\mathbf{x})$ (q instead of p), where $q(\mathbf{x})$ represents an unbiased sample from the world, rather than a selection made by the algorithm. With the dependency, the algorithm at iteration $n+1$ sees input $\mathbf{x}_{n+1}$ which is generated from a mixture between the objective distribution $q(\mathbf{x})$ and another distribution that captures the dependency upon the previous hypothesis h_n which biases the future inputs seen by the user.



(a)



(b)

Figure 2: In iterated learning, information is passed through selected data, (a) where the inputs, x, are independent of the inferred hypothesis, and (b) where the next inputs are selected based on the previous hypothesis. The latter case is more consistent with recommender systems and information filtering circumstances.

5. CONCLUSION AND FUTURE OUTLOOK

Recommender systems, and machine learning more broadly, have benefited greatly from benchmark datasets. However, recent successes in the field, and specifically applications "in the wild" have led to prolific use of data that are not curated in the traditional sense, but instead arise from human-algorithm interaction. This cyclical flow is rarely taken into account in algorithm design and analysis; nor is its impact (or dependence) on algorithms and humans well understood. We propose that the next generation of recommender systems be trained not on benchmark data, but on benchmark models. This approach is necessarily interdisciplinary and involves building on cognitive science research to develop models of human behavior that formalize the non-stationarity and non-independence of human behavior that cannot be easily captured in data alone.

Key to our approach is the focus on the sources and consequences of bias in data collected "in the wild". The two primary sources of bias are from algorithms and from humans. Recommender systems filter information with the goal of presenting humans with the most preferred content. This aspect of recommendation raises questions regarding the non-independence of choice behavior that have been documented in cognitive science for decades. Our approach suggests that it may be important to not only study people's choices, but also their beliefs about the domain in question and about the algorithm itself. The former has been a standard question in both human and machine learning for decades. The latter speaks to ways in which algorithms may not merely serve human goals, but that those goals may also be a consideration of the algorithm. We emphasize that

even user-tests do not make up for the gap in benchmarking validity because they typically focus on user satisfaction or engagement with what the algorithm suggested, and thus ignore the human cognitive aspect. When algorithms are able to both recommend content *and* reason about when and why people prefer something, we will be moving toward truly *human-recommender systems*.

6. REFERENCES

[1] F. G. Ashby, L. A. Alfonso-Reese, and E. M. Waldron. a neuropsychological theory of multiple systems in category learning. *Psychological Review*, 105:442–481, 1998.

[2] F. G. Ashby, W. T. Maddox, and C. J. Bohil. Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, 30:666–677, 2002.

[3] F. G. Ashby, S. Queller, and P. T. Berretty. On the dominance of unidimensional rules in unsupervised categorization. *Perception and Psychophysics*, 61:1178–1199, 1999.

[4] C. L. Baker, R. Saxe, and J. B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113:329–349, 2009.

[5] C. L. Baker, R. Saxe, and J. B. Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*, 2011.

[6] C. Basu, H. Hirsh, W. Cohen, et al. Recommendation as classification: Using social and content-based information in recommendation. In *AAAI/IAAI*, pages 714–720, 1998.

[7] A. Beppu and T. L. Griffiths. Iterated learning and the cultural ratchet. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2009.

[8] C. M. . Bishop. Pattern recognition and machine learning. *Springer*, 4(4), 2006.

[9] J. Bruner. The art of discovery. *Harvard Educational Review*, 31:21–32, 1961.

[10] J. R. Bruner, J. J. Goodnow, and G. A. Austin. *A study of thinking*. Wiley, New York, 1956.

[11] J. C. Caicedo, J. BenAbdallah, F. A. González, and O. Nasraoui. Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. *Neurocomputing*, 76(1):50–60, 2012.

[12] B. S. Eaves and P. Shafto. Unifying pedagogical reasoning and epistemic trust. In F. Xu and T. Kushnir, editors, *Advances in Child Development and Behavior, Volume, 43*, pages 295–319. Elsevier, San Diego, 2012.

[13] N. D. Goodman, C. L. Baker, E. B. Bonawitz, V. K. Mansinghka, A. Gopnik, H. Wellman, L. E. Schulz, and J. B. Tenenbaum. Intuitive theories of mind: A rational approach to false belief. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*, 2006.

[14] T. L. Griffiths and M. L. Kalish. A Bayesian view of language evolution by iterated learning. *Cognitive Science*, 31:441–480, 2007.

[15] S. Kirby. Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation*, 5:102–110, 2001.

[16] S. Kirby, M. Dowman, and T. L. Griffiths. Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104:5241–5245, 2007.

[17] Y. LeCun, C. Cortes, and C. J. Burges. The mnist database of handwritten digits, 1998.

[18] M. Lichman. UCI machine learning repository, 2013.

[19] S. B. Y. J. Linden, G. Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, 2003.

[20] R. D. Luce. *Individual choice behavior*. John Wiley, New York, 1959.

[21] D. Markant and T. Gureckis. Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, 143(1):94–122, 2014.

[22] D. McFadden. Quantal choice analysis: A survey. *Annals of Economic and Social Measurement*, 5:363–390, 1977.

[23] R. S. Michalski. Pattern recognition as rule-guided inductive inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2(4):349–361, July 1980.

[24] R. S. Michalski, I. Bratko, and A. Bratko. Machine learning and data mining; methods and applications. 1998.

[25] T. M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[26] O. Nasraoui, J. Cerwinske, C. Rojas, and F. A. González. Performance of recommendation systems in dynamic streaming environments. In *SIAM Data Mining (SDM)*, pages 569–574. SIAM, 2007.

[27] P. Shafto and N. D. Goodman. Teaching games: Statistical sampling assumptions for pedagogical situations. In *Proceedings of the 30th annual conference of the Cognitive Science Society*, 2008.

[28] P. Shafto, N. D. Goodman, and M. C. Frank. Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, 7:341–351, 2012.

[29] I. Simonson and A. Tversky. Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*, 29:281–295, 1992.

[30] K. Sparck Jones. Some thoughts on classification for retrieval. *Journal of Documentation*, 26(2):89–101, 1970.

[31] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT Press, Cambridge, MA, 1998.

[32] M. Tomasello. *The Cultural Origins of Human Cognition*. Harvard University Press, 1999.

[33] K. Train. *Discrete choice models with simulation*. Cambridge University Press, Cambridge, 2003.

[34] A. Tversky. Elimination by aspects: A theory of choice. *Psychological Review*, 79:281–299, 1972.

[35] J. I. Yellot. The relationship between luce's choice axiom, thurstone's theory of comparative judgement, and the double exponential distribution. *Journal of Mathematical Psychology*, 15:109–144, 1977.