Ethan Widener

3 May 2022

Data Collection for Recommendation Systems: Knowing is Half the Battle

In the 21st century, there is no commodity more valuable than data. The process of collecting and interpreting data underlies the activities and decisions of every contemporary corporation, government, and service provider. However, much of this data is only useful when fed into machine learning algorithms, such as recommendation systems, that can detect nuanced patterns and provide meaningful insights otherwise imperceptible due to the sheer volume and complexity of the data collected with modern infrastructure. Recommendation systems thrive on the availability of user data and thus, “the more you teach [them],” either explicitly or implicitly, “the better [they] can serve you—or manipulate you” (Domingos 264). These systems can use user data to construct digital profiles based on user interests, behavior, and demographics to streamline, or, without sufficient exercise of caution, usurp the decision-making process from users’ everyday lives. In contemporary society, data mining and influence from recommendation systems are inescapable. Therefore, it is essential to be aware of how user data is being collected and applied to power recommendation systems so that users can make informed decisions about the services they engage with, employing them to serve rather than manipulate.

It is no secret that corporations and organizations collect, analyze, and use the data that users provide them, but what remains esoteric to most is how these entities are collecting user data and what precisely they are doing with it. Kevin Roose, the author of *Futureproof*, a book intended to expose how machine learning is being used to influence our daily lives, writes that “today’s tech companies” are able to “generate detailed models of user behavior, and machine learning techniques…” that “let them discover patterns in enormous data sets” (85). These machine learning techniques, however, would be rendered ineffective without the data collection methods developed by parties interested in user data. According to Amy Webb, “the most challenging part of building a new AI system isn’t the algorithms… but rather collecting the right data” (143). While there is an abundance of data that can be collected, corporations and organizations are only interested in data they can apply to their models to generate a more complete picture of who their users are and how they behave. These processes of extracting relevant data from a user are known as feedback techniques, of which there are three primary approaches: explicit feedback, implicit feedback, and hybrid feedback.

Explicit feedback techniques are perhaps the most familiar to the average user. Explicit feedback techniques directly query the user for feedback on a product or service, asking them to “[assign] either numeric or score ratings for evaluating the product” or service (Das 8). These ratings “permit these judgments to be handled statistically to give distributions” that are then used by organizations to adjust their products or services and better tailor them to their userbase (Das 8). Explicit feedback also includes text-based responses that are provided by users but cannot be easily converted to a numeric scale. The advantages of this technique are that feedback provided by the user can be quantified as positive or negative, which “helps the user to specify what they like and what they don’t,” in turn providing the organization with a clear indication of user satisfaction with “accuracy” that “seems to be higher than” implicit feedback techniques (Das 8). However, explicit feedback techniques suffer from a dependence on context specific to each user that can be difficult to interpret from a numeric score or be extracted from text. This dependency leads to inconsistency in user ratings and may or may not accurately reflect a userbase’s “true opinion,” making explicit feedback techniques alone generally insufficient for canvasing a userbase or individual accurately enough for an effective recommendation system (Das 8).

While explicit feedback techniques receive data directly from the user, implicit feedback techniques analyze user behavior to extract relevant information without the user’s direct awareness. Cheng states that “it is particularly important” to collect implicit data because “it is necessary to make a detailed record of users’ visits to relevant web pages… and then make a summary” that can be digested by a recommendation system (71). Sources of implicit data include “browsing history, web consumption history, and mouse movements or even search patterns” (Das 8). While this data can be difficult to interpret, it can provide insights into a user’s behavior and preferences that explicit feedback cannot capture. Due to this type of data often being collected without a user’s direct knowledge, it can glean information from them that they might not otherwise provide directly. Implicit data can be invaluable to organizations looking to build a more robust profile for a user; however, it also raises privacy concerns, as users may not want recommendation systems to elicit certain information from them without their direct consent.

Due to the imperfect nature of both explicit and implicit feedback techniques, organizations often utilize a combination of both approaches, known as hybrid feedback, to feed their recommendation systems. The direct data gathered by explicit feedback techniques can be supplemented and reinforced by implicit data to generate a more concrete profile of a user so that a recommendation system can get “more value out of the data” and return “some of that value… to you in the form of more relevant ads and better service” (Domingos 271). This exchange of data for service is, as Domingos notes, “the bargain you make when you use Facebook” or any other online service (271). Modern corporations and organizations depend on the data that users provide, whether explicitly or implicitly. It is therefore imperative for users to understand how their data is being collected and how it is being used to influence their decisions. Because “corporations have a vastly greater ability to gather and use data than individuals,” this disparity “leads to an asymmetry in power” that can only be mitigated with an awareness of the specifics of the practice, allowing individuals to resist and compensate for influence from the organizations that each have a “sliver” of their data (Domingos 272, 275).

Data collection in a vacuum is not significant enough to influence the decisions or behavior of users of a service. Only when paired with a machine learning recommendation system designed to profile an individual from an incomplete picture does data begin to hold sway over a user’s decision-making process, making an understanding of how recommendation systems skew a user’s choices crucial. According to Roose and Da’u, “the world runs on recommendation engines” that “are utilized for solving information overload problems in areas such as e-commerce, entertainment, and social media” (Roose 84; Da’u 2709). Often, these recommendation systems are tailored to an individual’s data to provide the most relevant suggestions to that individual “based on their past behavior” (Das 7).

There is a myriad of models used by organizations to make recommendations, with the “most widely used” method being collaborative filtering, which can be further divided into user-based and item-based approaches (Da’u). In a user-based approach, the previous behavior of users is analyzed and cross-referenced “to calculate the item or else rating that the user may perhaps be interested in” (Das 7). In contrast, an item-based approach to collaborative filtering homes in on the similarities between entities that the user rates to recommend new items or services. With either approach, there is “collaboration” between entities that generates a web of recommendations and selects the most relevant recommendation for each user. Collaborative filtering, therefore, can affect future users based just on the behavior of past ones. Collaborative filtering also suffers from the cold start problem, a flaw in recommendation systems where inferences cannot be drawn for a user or item without sufficient data. Content-based filtering, in contrast, does not suffer from the cold start problem. In content-based filtering, “algorithms aim to suggest items or products which are alike to the items” that the user is currently or has previously viewed based on content descriptions (Das 7). It is distinguished from item-based collaborative filtering in that it does not require user ratings, which also prevents this approach from estimating the “excellence” of a product (Das 7). While this approach only considers data extracted about the item, in demographic filtering, only user data “like age, gender,” and “employment status” are considered in order to correlate a specific demographic that the user matches to products or services that demographic tends to use (Das 7). This type of recommendation system can often be controversial, as there are many privacy concerns associated with gathering and using this type of data, whereas the data used by other models are often expected to be collected, such as with knowledge-based recommendation systems. These systems explicitly ask the user for information about their preferences before generating recommendations based on the context provided. Often, corporations and organizations will use a combination of these models, called a hybrid, to cover their weaknesses while maximizing their advantages. Amazon, for instance, learns “customer preferences from their browse and purchase history,” as well as the products they and others have reviewed, consolidating this information into a “Bayesian approach” that “effectively deals with the issues arising from noise and sparsity in the data” (Rastogi).

Recommendation systems are powerful tools that serve as the engines driving contemporary decision making and can, through “mining our personal data for insights, … result in all of us living healthier, happier lives” (Webb 73). However, “today’s recommender algorithms are so powerful, and so deeply embedded in our systems, that they often function more like decider algorithms,” influencing users’ decisions and behavior in profound ways that they are often not aware of (Roose 90). Therefore, it is essential for users to have an awareness and understanding of how their data is being collected and how recommendation systems are operating so that they can compensate for this influence, which Rosse calls “machine drift,” and remain in control of their own decisions (Roose 81). Users should be aware enough to “tell [recommendation systems] as much as [they] want about [themselves],” and not just have them “learn indirectly from what [they] do” (Domingos 267). In an absence of this awareness, it is difficult to resist machine drift and utilize recommendation algorithms for a user’s advantage rather than their manipulation due to the omnipresence of these algorithms in modern life.

Recommendation systems, in their ubiquitousness, have become integral to contemporary society by driving the decisions of individuals and corporations alike. The data that these corporations and organizations collect, whether explicitly or implicitly, is used to tailor suggestions from imperfect models for individuals based on incomplete digital profiles. Without an awareness of what data is being collected and how it is being used in these systems, recommendation systems can usurp rather than facilitate an individual’s decision-making process. With a proper understanding of common data collection and recommendation techniques, however, users of any online service are able to reap the benefits of recommendation systems while avoiding their pitfalls, potentially improving their lives in a manner not possible without the utilization of all of the data they provide.

Works Cited

Cheng, TianMin. “Product Recommendation System Design.” 2nd international Conference on Information Management and Management Sciences, August 2019. *ACM Digital Library*. <https://dl.acm.org/doi/10.1145/3357292.3357314>.

Das, Debashis, et al. “A Survey on Recommendation System.” International Journal of Computer Applications. vol. 160, no. 7, 2017. *Researchgate*. 10.5120/ijca2017913081.

Da’u, Aminu, and Salim, Naomie. “Recommendation system based on deep learning methods: a systematic review and new directions.” vol. 53, 2020, pp. 2709-2748. <https://doi.org/10.1007/s10462-019-09744-1>.

Domingos, Pedro. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books, 2015.

Rastogi, Rajeev. “Machine Learning @ Amazon.” ACM SIGIR Conference, July 2018, Ann Arbor, MI. *ACM Digital Library*. <https://dl.acm.org/doi/10.1145/3209978.3210211>.

Roose, Kevin. *Futureproof: 9 Rules for Humans in the Age of Automation*. Random House, 2021.

Webb, Amy. *The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity*. PublicAffairs, 2019.