Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
AAAI Technical Report CR-13-01

Using Human and Machine Processing in Recommendation Systems

Eric Colson
Stitch Fix, Inc.
ecolson@stitchfix.com

Abstract

This paper presents a case study of an online apparel retailer called Stitch Fix, which uses both machine and human processing in its recommendation system.

Introduction

Customer-item recommendation systems are used by many ecommerce companies. For this task, machine-learning algorithms are efficient at processing *structured* data for ranking vast catalogs of merchandise in the context of a customer. Yet, machines are notoriously challenged when it comes to processing *unstructured* data such as images and free-form text. Humans, on the other hand, can process unstructured data effectively. They can also better contextualize the results and perceive more nuanced distinctions vs. machines alone. However, human processing is inefficient on large sets of unranked items. By using the two resources together in a single system, more data and processing can be leveraged.

About Stitch Fix

Stitch Fix is a personalized styling service. It offers apparel items over the internet, similar to an ecommerce company, but with the all-important difference that the customer does not pick out the merchandise herself. Instead, an algorithm will select a set of five pieces of merchandise on behalf of the customer and send them directly to her, sight unseen. Once received, the customer evaluates the merchandise and can send any or all of it back with free shipping both ways. She only pays for the merchandise she decides to keep.

Expert Humans for Latent Data Capture

Standard data that is used to describe apparel items can often fail to distinguish them. For example, two items may have the same brand, size, color, and price, yet have very different sales rates or might even be purchased by different types of customers. However, professional stylists and fashion merchandisers can often make important distinctions that explain the variation. They may notice the subtleties of the cut and how it makes the item fit loosely on the shoulders of a small-framed customer. Or, they may spot how the stitching pattern on a skirt transforms the piece from being classic in style to edgy. These distinctions can be significant features for determining an item's relevancy to a specific customer. Yet, they are latent to all but the fashion expert. But, by persisting them into data, these distinctions can be captured and made available as features for machine learning. Our system provides a human-machine interface to capture these latent attributes. To mitigate subjectivity, classifications from multiple experts are aggregated (Parameswaran et al. 2012). The attributes can be of any data type including free-form text and images.

The system also includes an interface to collect data that describes the customer. The customer herself provides this data, however, the expert determines what information to collect. This includes the counterparts to the distinctions made on the merchandise. For example, the system collects data on the customer's preference for fit on the shoulders and whether or not edgy clothes are desired. Other data attributes collected include: height, weight, age, favorite brands, images of the customer, links to the customer's Pinterest pages, request notes …etc. Like the attributes that describe the merchandise, the customer attributes can be of any data type.

Leveraging Human and Machine Processing

The merchandise and customer data, along with past interaction data, represents the corpus of knowledge available for processing in order to determine relevancy. The data can be divided into two classes by format. Data such as size, age, preference for fit on the shoulders, ...etc. are structured and represented by well-known domains of values (e.g. "XL", "34", "loose"...etc.). Other data such as images of the merchandise or customer, request notes, pins from Pinterests, ...etc. are unstructured; they require interpretation to extract the meaning they convey. To exhaust the full corpus of knowledge – the structured and

the unstructured data - our system leverages both machine and expert-human processing in its algorithm.

Machine processing is used to efficiently perform computations involving complex analytic logic on large data sets (e.g. logistic regression, support vector machines, random forest ...etc.). However, only the structured data is considered. The unstructured data is deferred for consideration by human processors. The output of the machine processing is a relevancy score for each item (i.e. the probability the customer will like the item). This score will be used by humans in the subsequent process.

Expert-human processing is used to perform several tasks deemed outside the purview of machines. First, humans are capable of processing the unstructured data, which can be used to further contextualize the relevancy between the customer and the merchandise. For example, the free-form text of a request note may read, "*I need clothes for an upcoming cruise*". Or, by viewing the customer's Pinterest page, a human stylist can get a sense of her preferences – even those not explicitly stated. This information changes the context for the recommendations materially and expert humans can then select items accordingly from the ranked output of the machines.

Expert-human processing is also used to curate the selections. While machines do their processing under the assumption of item-independence, human processing is used to synthesize concepts from the individual items to create a curated set that collectively represent better relevance vs. the sum of the individual relevancy scores.

Coordinating Resources

To coordinate the processing across human and machine resources, a queuing system is used. When a customer request is received, it is routed to a queue for machine processing of the structured data. Once the machine-generated results are available, they are routed to an internal crowdsourcing queue consisting of dozens of geographically distributed human stylists. The logic for determining which stylist the results are routed to can be tuned using several parameters. The stylist accesses the queue and the associated results through a graphical user interface. It allows her to incorporate unstructured information by displaying images and textual descriptions of the merchandise and the customer. She can then review the recommendations, curating and making adjustments as necessary. Once the selections are finalized, the request is routed to logistics for a pick-pack-and-ship process. Currently, the vast majority of customer requests are processed by just a single stylist (multiple stylists are used only during a new stylist's evaluation period). In the future, we will experiment with systematically aggregating the judgments from multiple stylists in order to improve recommendations (Parameswaran and Polyzotis 2011). For now, recommendation quality is evaluated on various ex-post metrics (Garcia-Molina and Koutrika 2011).

This sequencing of the resources is important. Human processing is slower (measured in minutes) relative to that of machines (measured in milliseconds). Since the machine-generated recommendations are produced first and the output ordered, the human stylists only needs to be presented with the most relevant merchandise qualified on any number of criteria (e.g. *return the best cruise attire in descending order of relevancy*). Hence, the amount of merchandise the humans needs to process is minimized.

The two resources combined exhaust more data (the structured and the unstructured) and more processing capability (complex computation and expert judgment).

Improvement Over Time

Our use of both human and machine resources within a single system enables improvement over time. That is, the system gets better due to the sharing of feedback and knowledge between the resources and this is separate from any improvement owing to intra-process optimization. The selections made by expert humans provide feedback to the machine ranking process. This can be used to improve the accuracy of the machines and/or to expose biases on the part of the humans. In some cases, the explanations for the discrepancy between machine rankings and human selections can be captured and persisted as new features. With each new feature persisted, the corpus of knowledge available to the broader system is increased.

In addition, human oversight enables faster development of machine algorithms. Recommendations made by machines can sometimes be highly relevant, yet also inappropriate (e.g. failing to recognize a past purchase as gift for someone else). These tend to be edge cases, but anticipating them can be expensive. The human processors in our system catch the majority of these cases before they are surfaced to customers. With edge-cases covered, the machines processing can evolve at a faster pace.

References

A. Parameswaran, H. Garcia-Molina, H. Park, N. Polyzotis, A. Ramesh, and J. Widom. Crowdscreen: Algorithms for Filtering Data with Humans. ACM SIGMOD International Conference on Management of Data, Scottsdale, Arizona, May 2012.

A. Parameswaran and N. Polyzotis. Answering Queries using Humans, Algorithms and Databases. Conference on Innovative Database Research (CIDR), Asilomar, USA, Jan 2011.

H. Garcia-Molina, G. Koutrika, and A. Parameswaran. Information Seeking: Convergence of Search, Recommendations and Advertising. Communications of the ACM, Viewpoint Article, Nov 2011.