Report date 11/01/2025

Paper report:

**Overview:**

Snapcase is a system to apply low-latency unlearning on a production size dataset of user purchases. The dataset contains 33Mln purchases and 200K users.

The system is based on:

* Data stored in a relational database
* Data preparation pipeline implemented with Differential Dataflow
* The first part of TIFU (the NBR algo used) is implemented in Differential Dataflow
* kNN top-k neighbours stored in Caboose (custom data structure that allows unlearning)
* Dynamic computation of the actual recommendation based on the own user repeated purchase history and on the other user (collaborative) purchase component

The paper is interesting since it shows an end to end system that allows unlearning (low-latency unlearning).

The relevant components are Differential Dataflow (framework for IVM pipelines) and the modified data structure to run the NBR (Next Basket Recommendation) algorithm on (TIFU-kNN namely) while allowing unlearning on it.

The first part of the paper describes how important unlearning and low-latency unlearning are.

NBR and TIFU are then explained and Caboose is introduced to allow TIFU unlearning on the top-k closest neighbours. Differential Dataflow is introduced which also partially implements TIFU, namely the embeddings computation.

**Experimental metrics and scenarios:**

The goal is to unlearn products from customer recommendation based on ethically motivated changes in the customer shopping habits regarding: alcohol addiction, obesity, high carbon footprint. We want not to suggest wine to a client that's dealing with an alcohol addiction.

Instacart dataset contains:

* 33M purchases
* 206K users
* 3.4M baskets
* 49K distinct items

The user can remove the unwanted product and the graph is updated and also the final recommendation probabilities are updated.

**Repo:**

<https://github.com/deem-data/snapcase>

**Authors:**

**BIFOLD**

Sebastian and Stefan et alii

**Year:**

2024