# Towards a Privacy-aware Recommendation System for Manga Reading Application

Mhd Irvan[1,a)] Toshiyuki Nakata[1] Rie Shigetomi Yamaguchi[1]

**Abstract**: Reading habits can potentially reveal many characteristics of the readers that they likely want to keep private to themselves. Furthermore, many on-demand reading applications nowadays are being deployed as smartphone applications, allowing even more delicate data to be detected and shared away from the phone itself. These information are useful to feed into centralized machine learning programs to, for example, recommend interesting contents. This paper argues that it is possible to build reliable recommendation systems without gathering those data into a centralized place, beyond the users' control. We propose a privacy-preserving machine learning approach that can be applied to recommendation systems. This approach is tested on a manga reading application dataset to demonstrate its usefulness in real world usage.

**Keywords**: Recommendation System, Privacy, Machine Learning

## 1. Introduction

Recommendation systems are reliable tools to keep users continuously engaged [1]. The personalized experience offered by those systems helps users to find contents that are particularly interesting for them. Recommendation systems achieve accuracy by relying on personal information and user behaviors [2]. However, this also means that these systems have the potential to breach user privacy and may infer sensitive information [3].

Furthermore, the rapid adoption of mobile devices such as smartphones into people's daily life implies that even more delicate data can potentially be detected by the devices' sensors [4]. Mobile applications, such as media-streaming services, track various user behaviors during their interaction with the application and gather those data in a centralized place away from user's device to train recommendation systems at a remote server [2]. These behaviors, such as viewing time or reading habit may reveal sensitive information about a user's lifestyle and personal life [5].

1. Graduate School of Information Science and Technology The University of Tokyo
a) irvan@yamagula.ic.i.u-tokyo.ac.jp

In this study, we explore the possibilities of making recommendation systems to be privacy-aware and propose a method to develop such systems by keeping the data local at user's device and not gathered in a centralized place. We train our proposed method with a data set provided by Shogakukan Inc, containing information about users' interaction with one of their smartphone applications called Manga One. We then demonstrate how a recommendation system can be build on top of this type of data.

## 2. Related Works

Privacy issues in machine learning, such as those in recommendation systems have become important points of discussion recently. Boutet et al [6] proposed an approach to obfuscate user profiles to achieve privacy. This prevent a complete profile of users to be shared with recommendation systems.

Basha et al [7] proposed an encryption method to encrypt user private data. The server tasked for generating recommendation is then allowed to perform certain types of limited computations on the ciphertexts. This way, certain information privacy can be preserved.

Puglisi et al [8] proposed data perturbation methods for their recommendation systems. By perturbing data, noises

are injected to users' private data before sending them to the server to generate recommendations. These noises help covering sensitive information of users.

These data obfuscation, encryption, and perturbation are reliable approaches in protecting uses' privacy. However, they still require the data itself to be gathered at a remote server, away from user's device. Remote server are prone to breach, and while the stored data may not be in complete forms, possibilities exist where private information can be inferred even from from incomplete data [5] [9].

## 3. Proposed Approach

Here, we propose a method to generate recommendation without gathering users' data in a centralized place, by keeping the data local on the users' device. Our proposed method is inspired by Federating Learning [10] concepts to distribute the machine learning process into the users' device itself, instead of learning at a remote server (Fig. 1).

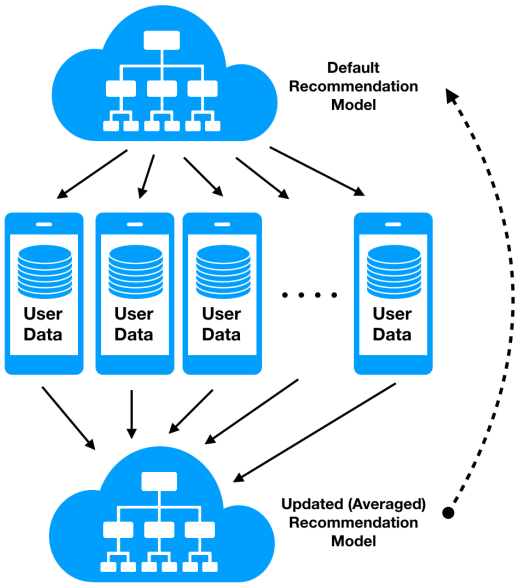


Figure 1: Distributing recommendation model across users' device and periodically updating the model

Initially, a default recommendation model is generated at a server. Next, this model is copied across users' device to interact with the user data locally. Throughout the learning process, the parameters of the recommendation model is updated based on the user's feedback. Different devices will update these parameters differently according the user's preference. After several updates, all parameters across devices are averaged to create an updated default recommendation model. This way, since no user data are shared outside the device itself, privacy can be kept locally on the device.

For the recommendation model itself, we implemented a Distributed Classifier System [11] that was developed as a machine learning approach for smart home. This model has been found to be able to successfully learn the preferences of smart home residents [11], and was also previously implemented as a recommendation system for educational services [12].

This classifier system is essentially a population of condition-action strings that represent an action to recommend given certain conditions are met. This population is updated through cycles of Genetic Algorithms and Reinforcement Learning [11] (Fig. 2)

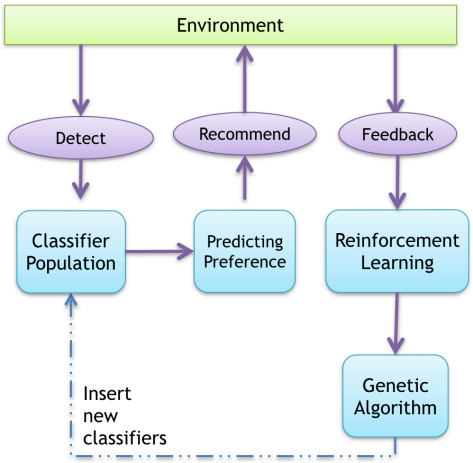


Figure 2: Architecture of Proposed Classifier System

To recommend, the system first calculates the prediction array *P(A)* for classifiers *C* that matched with the current conditions, referred as matched set *[M]* collection, using this formula:

$$P(A) = \frac{\sum_{C.a=a \wedge C \in [M]} C.p \times C.f}{\sum_{C.a=a \wedge C \in [M]} C.f} \quad (1)$$

Essentially, P(A) reflects the average of prediction of classifiers in [M] that recommends *a*. The system then

recommends the action that maximizes the prediction array.

$$A_{\max} = \arg\ \max_A P(A) \tag{2}$$

As each device will evolve the classifiers in different ways based on the recommendations accuracy, parameters for all classifier systems across every users are averaged to create a new default classifier system to distribute. This new default classifier system is then again copied across all users' device to further enhance the recommendation accuracy.

## 4. Early Experiment

We conducted an early experiment to test the feasibility of our proposed approach. The experiment was done with a real-world dataset in a simulated environment.

### 4.1. Dataset

We tested our proposed approach with a dataset provided by Shogakukan Inc for one of their smartphone applications called Manga One. The data was generated through a collaborative research project between the company and The University of Tokyo.

Through announcements of the project inside Manga One Application (Fig. 3), a total of 49261 users participated in the program. The dataset contains reading data of Manga One users between August 24, 2015 to December 17, 2015.

There are various logs inside the data set, and here we consider the below log data for our recommendation system:

- Anonymized user information
- Title and chapter information
- Access information

These information are used to represent the chromosome string for the distributed classifier system's genetic algorithms.

Since this research is still in its early phase, to confirm the feasibility of our approach, in this paper we focused on the 100 most active users in the dataset. These users are the users with the largest reading list.


Figure 3: In-app announcement (in Japanese language) of a collaborative project between Manga One and The University of Tokyo (Top: iOS, Bottom: Android)

### 4.2. Early Simulation Result

For the simulation, we separate the log data of each user away from each other into their own simulated device. An initial classifier systems is generated and copied into each of those simulated devices.

As a measurement for the recommendation system, titles that are continuously read into more than 5 chapters by a user are considered to be interesting contents for them. Of all the interesting contents, 80% were addressed as a training set and 20% were addressed as a validation set for the classifier system on the respective simulated device. This 80:20 ratio are repeated 5 times with different dividing point each time, essentially generating a 5-fold cross validation process.

Table 1 summarizes the precision and recall value generated by the recommendation system. Our early results show that the proposed approach of this research reliably recommend interesting contents for the user more often than not. Although it is not uncommon for non-privacy approach of recommendation system to produce higher accuracy, the privacy benefits are arguably big. We believe this initial results show a promising approach and further optimization techniques can be implemented to potentially minimize the loss of accuracy for the price of privacy.

Table 1: Precision and recall values at each fold

| | Precision | Recall |
|---|---|---|
| **Fold-1** | 65.7% | 70.4% |
| **Fold-2** | 63.4% | 68.4% |
| **Fold-3** | 64.8% | 69.2% |
| **Fold-4** | 65.3% | 70.1% |
| **Fold-5** | 63.6% | 68.5% |

## 5. Conclusion & Future Works

In this paper, we proposed a model for privacy-aware recommendation system. Privacy is achieved by preventing user data from leaving user's device. Instead of sending data into a remote centralized server, we suggested sending recommendation model into user devices, inspired by federated learning concepts. Our implementation of distributed classifier system showed that privacy-aware, distributed recommendation systems are feasible approaches with reliable accuracy. We currently plan to scale our model to work with the whole contents of Manga One dataset. Federating models other than classifier system models are also being considered.

## References


[1] Davidson, J., et al.. The Youtube video recommendation system. Proceedings of the fourth ACM conference on Recommender Systems, 2010, p.293-2996.

[2] Lu, J., et al.. Recommender system application developments: A survey. Decision Support Systems, Elsevier, 2015, Vol. 74, p.12-32.

[3] Friedman, A., et al.. Privacy Aspects of Recommender Systems. Recommender Systems Handbook, Springer, 2015, p. 649-688.

[4] Hassan, M.M., et al.. A robust human activity recognition system using smartphone sensors and deep learning. Future Generation Computer Systems, Elsevier, 2018, Vol. 81, p307-313

[5] Salamatian, S., et al.. Managing Your Private and Public Data: Bringing Down Inference Attacks Against Your Privacy. IEEE Journal of Selected Topics in Signal Processing, 2015, vol. 9, p1240-1255

[6] Bought, A., et al.. Privacy-preserving distributed collaborative filtering. Computing, Springer, 2016, Vol. 98, Issue 8, p.827-846

[7] Basha, S., et al.. A Practical Privacy-Preserving Recommender System. Data Science and Engineering, Springer, 2016, Vol. 1, Issue 3, p.161-177

[8] Puglisi, S., et al.. On Content-based recommendation and user privacy in social-tagging systems. Computer Standards & Interfaces, Elsevier, 2015, Vol. 41, p.17-27

[9] Okkalioglu, B.D., et al.. A survey: deriving private information from perturbed data. Artificial Intelligence Review, Springer, 2015, Vol. 44, Issue 4, p.547-569

[10] Smith, V., et al.. Federated Multi-Task Learning. Proceedings of Neural Information Systems, 2017

[11] Irvan, M., and Terano, T.. Distributed Classifier System for Smart Home's Machine Learning. Agent-based Approaches in Economics and Social Complex Systems IX, Springer, 2017, p. 191-197

[12] Irvan, M., and Terano, T.. Group Recommendation System for E-Learning Communities: A Multi-agent Approach. Advances in Social Computing and Digital Education, Springer, 2016, p. 35-46