The Web Alter-Ego project

Rachid Guerraoui (EPFL) & Anne-Marie Kermarrec (Inria)

Google Focused Award
Personalization is now ubiquitous
Why is personalization challenging?

• **Huge volume** of data: small portion of interest
• Dynamic and diverse interests
• Interesting stuff does not come always from friends
• Classical notification systems do not filter enough or too much

KNN-based collaborative filtering
The Web-Alter ego project

Extracting like-minded Internet users should be a basic Web service

Goals of Web Alter-Ego: cross-apps KNN-based collaborative filtering

1. Provides an efficient scalable infrastructure
2. Provides privacy guarantees

TEAM
Nitin Chiluka (postdoc Inria)
Nupur Mittal (PhD student Inria)
Rhicheek Patra (PhD student EPFL)
Antoine Rault (PhD student Inria)
Masha Taziki (PhD student EPFL)
Jingjing Wang (PhD student EPFL)
Main results so far


HyRec: Leveraging Browsers for Scalable Recommenders

Antoine Boutet, Davide Frey, Rachid Guerraoui, Anne-Marie Kermarrec, Rhicheek Patra
Middleware 2014
Personalization

Personalization schemes are resource greedy

- Fully decentralized systems, scalable but difficult to manage
- Centralized systems need huge computational power

Democratizing personalization is also crucial for small web content providers
HyRec’s challenge

Traditional centralized architecture

HyRec architecture

Online recommendation (front-end server)

Data

Offline Knn selection (back-end servers)

Data

HyRec Server (front-end server)

Personalization job
**HyRec: tasks to offload**

**User machine**
- Recommendation
- KNN selection

**Browser**

1: Client request

2: Candidate set

3: Update KNN

**Server**
- Global data structure
- Profile table
- KNN table

- Personalization orchestrator

**Sample:** Identify the candidate set (Two-hop neighborhood + k random)

**Orchestrator:**
- Personalization job (json) containing profile + profiles of users in the CS
- Update the knn table

No data stored at the client

Javascript (Interaction with the server's api)

- KNN computation
- Compute recommendations
View similarity

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Users</th>
<th>Items</th>
<th>Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>MovieLens1</td>
<td>943</td>
<td>1700</td>
<td>100,000</td>
</tr>
<tr>
<td>MovieLens2</td>
<td>6,040</td>
<td>4000</td>
<td>1,000,000</td>
</tr>
<tr>
<td>MovieLens3</td>
<td>69,878</td>
<td>10,000</td>
<td>10,000,000</td>
</tr>
<tr>
<td>Digg</td>
<td>59,167</td>
<td>7724</td>
<td>782,807</td>
</tr>
</tbody>
</table>

HyRec remains within 20% of the ideal KNN
Recommendation quality

Less than 13% below the best case

- HyRec
- Exhaustive p=24h
- Exhaustive p=1h
- Exhaustive best

NB of recommendations
HyRec versus the client load

Impact of HyRec

Negligible disruption of HyRec

Impact of the client load

50% load
<60ms on smartphone
<10ms on laptop
HyRec versus a centralized recommender

Impact of the profile size

Impact of the number of requests
Take away message

Scalable recommendation engines

Decentralized algorithms design

Hybrid infrastructures
D2P: Distance-Based Differential Privacy in Recommenders.
VLDB 2015
About privacy

Ex: Netflix challenge 2 and IMDB (Internet Movie Database)

« privacy expert Larry Ponemon says that Netflix could have likely avoided the matter altogether by using a technique called “data masking” that would have randomized its data set while still keeping the data relevant to developers »
Problem statement

1) Collaborative filtering relies on users profiles
2) Privacy guarantees needed

D2P: Distance-based Differential Privacy protocol: probabilistic substitution techniques to create the Alter-ego profile
Differential Privacy [Dwork 2006]

\[ \frac{\text{Prob}(Q(D))}{\text{Prob}(Q(D+/-1))} \leq e^\varepsilon \]

\[ \frac{\text{Prob}(R|\text{true world} = D)}{\text{Prob}(R|\text{true world} = D+/-1)} \leq e^\varepsilon \]

The released result \( R \) gives minimal evidence about whether or not any given individual contributed to the data set.

Adding (Laplacian) noise
DP2: DP applied to recommenders

- **DP**: Avoid any user to guess, based on her recommendations whether some other users has one item in her profile.

- **D2P**: And any item within some distance $\lambda$ from $I$.

D2P builds an alter-ego profile where some items are probabilistically replaced.
Technical challenge: trade-off

Distance to the original profile

Privacy

Quality
Example

D2P selects
- movies with distance less than an upper bound with prob. $p$,
- random movies with prob. $1-p$
D2P Recommender

1- A group $G_i$, contains all items with distance less than $\lambda$ from $i$

Distance between items $(i$ and $j) = (1/\cos_{sim}(i,j)) - 1$

2 - Create Alter-egos profile for each user (item substitution)

3 – KNN computation

4 – Recommendations
D2P Components

- **Selector**: This component *decides* whether to replace an item with a *close* item or *any* item.

- **Profiler**: This component builds the *Alter-Ego* profiles by *replacing* the items based on Selector’s decision.
Construction of the alter-ego profile

User Profile

Item x

Will be replaced by a group item

Item from G_x

1-p^*

1-p

p^*

p

Will be replaced by a random item

Random item

Alter-egosProfile
Distance-based Differential Privacy

For any two adjacent profile sets $D_1$ and $D_2$, where $U$ denotes any arbitrary user, $S$ denotes any possible subset of elements and $\text{GRP}(S)$ denotes union of element-wise groups of items in subset $S$, then any mechanism $R$ is private if the following inequality holds:

$$\frac{\Pr[R(D_1,U) \in \text{GRP}_\lambda(S)]}{\Pr[R(D_2,U) \in \text{GRP}_\lambda(S)]} \leq e^\varepsilon$$

We show (Theorem 1) that a mechanism $M$ relying on Alter-egos profile is an ($\varepsilon$, $\lambda$) mechanism.
Experimental evaluation
Experimental setup

• Training set (80%) – Test set (20%)

• Metrics
  • Precision = $\frac{T_p}{T_p+F_p}$
  • Recall = $\frac{T_p}{T_p+F_p}$

• Datasets
  • MovieLens (100k ratings, 943 users, 1602 movies)
  • Jester (4.1M ratings, 73 421 users, 100 jokes) – 500 users
Impact of Rating Density

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Users</th>
<th>#Items</th>
<th>Ratings</th>
<th>RD(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jester</td>
<td>500</td>
<td>100</td>
<td>36000</td>
<td>71.01</td>
</tr>
<tr>
<td>ML1</td>
<td>940</td>
<td>1680</td>
<td>99647</td>
<td>6.31</td>
</tr>
<tr>
<td>MLV₁</td>
<td>470</td>
<td>840</td>
<td>76196</td>
<td>19.3</td>
</tr>
<tr>
<td>MLV₂</td>
<td>470</td>
<td>840</td>
<td>16187</td>
<td>4.1</td>
</tr>
<tr>
<td>MLV₃</td>
<td>470</td>
<td>840</td>
<td>6317</td>
<td>1.6</td>
</tr>
<tr>
<td>MLV₄</td>
<td>470</td>
<td>840</td>
<td>750</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Effect of Selector probability $p$ (MovieLens)

The lower $p$ (fewer random substitutions) the better the recommendation quality
Effect of Selector Probability $p$ (jester)
Effect of Profiler Probability ($p^*$) (MovieLens)

The higher $p^*$ (the closer to the true profile) the better the recommendation quality

(a) Precision@N Comparison.  
(b) Recall@N Comparison.  
(c) Precision-Recall Comparison.
Effect of Profiler Probability $p^*$ (jester)
Overhead

- We compare the overhead of our system with the overhead in [1]

\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{Datasets} & \text{$D2P$ Overhead} & \text{$DP_\delta$ Overhead} \\
& RL & Online & Offline & Offline \\
\hline
ML1 & 196ms & 32ms & 4.54s & 120s \\
Jester & 24ms & 12ms & 162ms & 740ms \\
\hline
\end{array}
\]

To take away

Low-overhead solution
Extension of differential privacy to recommenders

Future plans in Web Alter-Ego

• Anonymous recommenders
• Quantifying the privacy impact of a click
• Impact of cross-applications
THANK YOU