Exploiting Personal Data, Preserving Privacy

Pradeep K. Murukannaiah and Munindar P. Singh

Department of Computer Science
North Carolina State University
Personal Data and Applications: Current Setting

Murukannaiah and Singh (NCSU)
Current Setting: Shortcomings

Potential for privacy loss

- Users’ personal data stored by several third-parties

Fragmented user experience

- Each application may interpret users’ data differently

Difficult to deliver a personalized experience

- Applications often rely on data aggregated from multiple users
The *middleware* maps raw sensor data into high-level contexts and shares contexts with applications, respecting user’s privacy policies.
Benefits of the Middleware Setting

Potential benefits to an end user

- Uniform location information architecture
  - Enhances learnability
- Simplified interaction design
  - Enhances ease of use
- Middleware-based service design
  - Preserves privacy

Potential benefits to a developer

- Separation of concerns
  - Simplifies development process
  - Yields modular implementations
The Disconnect between Data and Intent

An example: Intelligent Ringer application

- Automatically sets a ringer mode for incoming calls
- Consumes data from Bluetooth sensor

Why does the ringer application require from Bluetooth sensor?
Bridging Data and Intent

Xipho: A methodology for engineering personal agents

- Metamodel: Actor, goal, plan, belief, resource, and dependency
- Spans requirements, design, specification, and implementation phases
Bridging Data and Intent

Xipho maps *data abstractions* to *agent capabilities* via systematic steps

Example: \( \text{Activity} = ?A_1 \land \text{Social circle} = ?S_1 \land \text{Neighbor’s context} = ?N_1 \land \text{Caller’s context} = \text{Emergency} \rightarrow \text{Set as loud} \lor \text{Set as vibrate} \)
The middleware

- Elicits context instances
- Recognizes context instances
- Acquires context resources
Platys Middleware: Implementation

- Android Phone
  - API (Android IDL)
- Application Manager (Android Service)
- Tagging Subsystem (Android Activity)
- Shared Information Store (Dropbox Sync API)
  - Place Labels
  - Sensor Data
- Personal Computer
- Event & Action Logger (Android Service)
- Platys Reasoner (Apache Daemon)
  - Place Model (Semi-supervised)
Platys Reasoner

Platys Reasoner learns a user’s contexts from raw sensor data

**Traditional unsupervised approaches**

- Often require frequent sensing
- Do not capture subjective nuances

**Traditional supervised approaches**

- Several labels for each place to perform well

**Platys Reasoner: A middle ground solution**

- Active learning seeks to reduce labeling effort
- Semi-supervised learning deals with infrequent sensor data
Platys Reasoner: Intuition

Stream of intermittent sensor data and place labels

Active learning enable intelligent place labeling

Semi-supervised learning exploits latent structure
Platys Reasoner: Active and semi-supervised learner

- Pool of Unlabeled Sensor Data Instances
  - Active Learning
    - Prompt the User to Label A Few Labeled Instances
    - A Few Labeled Instances
      - Infer New Labels
        - Several Labeled Instances
          - Classify Instances
            - Filter Non-Places
              - Places from Sensor Data
  - Training
  - Semi-Supervised Learning
    - Infer New Labels
  - Iterative Clustering

Murukannaiah and Singh (NCSU)
Exploiting Personal Data, Preserving Privacy
Evaluation

Platys middleware

- Simplifies application development, enhancing privacy and user experience of applications produced

Xipho methodology

- Connects data and intent, enhancing software comprehensibility

Platys Reasoner

- Learns contextual abstractions from raw sensor data, reducing the user effort required for training the reasoner
Platys Developer Study: Design

Factor and Alternatives

- Framework: Platys middleware and Android location provider

Hypotheses

1. **Time and effort**: Platys middleware reduces time and effort required to implement an application compared to the control setting
2. **Code quality**: Platys middleware yields more modular implementations than the control setting
3. **Usability and privacy**: Platys middleware yields more usable and privacy-preserving applications than the control setting
Platys Developer Study: Design

Subjects

- 46 students from a graduate-level computer science course
- Balanced group for each alternative (education level, Android and general programming experience)

Study Unit

- An application to implement (Ringer Manager)
- Functional requirements specified as three–four use cases
- Usability requirement specified as broad-brush guidelines
Analyses

Modeling time (continuous value)

- Equality of mean values: Two-tailed $t$-test
- Equality of variance values: Two-tailed $F$-test
- Normality tests ($K$-$S$ test) conducted

Modeling effort, model completeness and comprehensibility
(ratings on a scale of 1–7)

- Equality of medians: Two-tailed Wilcoxon’s ranksum-test
- Normality assumption not required
Platys Developer Study: Results

\[ \mu_P \neq \mu_C \quad (p = 0.001) \]

\[ \mu_P < \mu_C \quad (p = 0.081) \]

\[ \mu_X \neq \mu_C \quad (p = 0.311) \]

\[ \mu_P < \mu_C \quad (p = 0.021) \]
Platys Developer Study: Results

\[ \mu_P < \mu_C \quad ** \quad (p = 0.001) \]

\[ \mu_P < \mu_C \quad ** \quad (p = 0.001) \]

\[ \mu_P < \mu_C \quad (p = 0.272) \]
Xipho Developer Study: Design

Factor

- Methodology

Alternatives

- Xipho and Tropos

Hypotheses

1. **Time and effort:** Xipho reduces modeling time and effort compared to Tropos
2. **Completeness:** Xipho models are more complete than Tropos models
3. **Comprehensibility:** Xipho models are easier to comprehend than Tropos models
Xipho Developer Study: Design

Subjects

▶ 46 students from a graduate-level computer science course
▶ Balanced group for each alternative (education level, modeling experience, and programming experience)

Study Unit

▶ Three applications to practice, model, and verify
▶ Requirements specified as three–four use cases

<table>
<thead>
<tr>
<th>Phase</th>
<th>Task</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Practice</td>
<td>Alarm</td>
<td>Motivator</td>
<td>Reminder</td>
</tr>
<tr>
<td>2</td>
<td>Modeling</td>
<td>Motivator</td>
<td>Reminder</td>
<td>Alarm</td>
</tr>
<tr>
<td>3</td>
<td>Verification</td>
<td>Reminder</td>
<td>Alarm</td>
<td>Motivator</td>
</tr>
</tbody>
</table>
Xipho Developer Study: Results

- **Practice hours per scenario**
  - $\mu_X < \mu_C$ ($p = 0.325$)
  - $\sigma^2_X < \sigma^2_C$ ($p = 0.087^*$)

- **Modeling hours per scenario**
  - $\mu_X < \mu_C$ ($p = 0.046^{**}$)
  - $\sigma^2_X < \sigma^2_C$ ($p = 0.563$)

- **Modeling effort (% responses)**
  - $\tilde{\chi}_X < \tilde{\chi}_C$ ($p = 0.01^{**}$)
Xipho Developer Study: Results

- Three subjects rated each model for completeness and comprehensibility.
- Ratings with insufficient inter-rater agreement (Krippendorff’s $\alpha < 0.67$) were discarded.

\[ \tilde{x}_X > \tilde{x}_C \quad (p = 0.206) \]

\[ \tilde{x}_X > \tilde{x}_C \quad (p = 0.029**) \]
Platys Reasoner User Study: Design

Data Collection
- 10 users employed an Android phone installed with Platys middleware as their primary phone for 10 weeks
- Platys middleware collected sensor readings
  - GPS, WiFi, Bluetooth sensor readings, and Google POI data
- Each user labeled places of his interest multiple times

Baselines
- Two staypoint-based approaches
- Two supervised approaches (logistic regression and SVM)
Platys Reasoner User Study: Results

Place Radius and Duration (log scale)

- **Platys (Place-or-not)**
- **Platys (Which-place)**
- **Staypoint [Zheng 2012]**

Staypoint for optimal place radius and duration for each user A–F

- 3 mins
- 30 mins
- 3 hours
- 1 day
- 20 m
- 200 m
- 1.2 km
- 96 km
Platys Reasoner User Study: Results

Number of labels trained with

- SVM
- SVM + Active learning

F-Measure (averaged across users) vs Number of labels trained with

Murukannaiah and Singh (NCSU) Exploiting Personal Data, Preserving Privacy
Platys Reasoner User Study: Results

![Graph showing F-Measure (averaged across users) vs. Number of labels trained with SVM + Active learning and Semi-supervised SVM + Active learning.]
Papers


Thank You