АРХИТЕКТУРА ПРОГРАММНОГО СРЕДСТВА ДЛЯ АНАЛИТИКИ И ВИЗУАЛИЗАЦИИ ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ

П.И. Банокин, Г.П. Цапко
Томский политехнический университет
pavel805@gmail.com

Введение

Данные о поведении пользователей программного обеспечения могут быть источником ценной информации, используемой для персонализированного маркетинга и рекламы, анализа и предсказания нагрузки на ПО и предотвращения внутренних утечек данных. Многие существующие решения в качестве источников данных о поведении пользователей используют низкоуровневую информацию: потоки сетевых данных или индивидуальные особенности использования клавиатуры [1] [2]. Высокоуровневые источники поведенческих данных (команды операционной системы, вид операции с данными и др.) позволяют проводить более точный анализ поведения пользователей [3]. Представленная программная система рассматривает поведение пользователя, как индивидуума, использующего набор программных приложений с разными учетными записями.

Компонентная архитектура программной системы

Программная система для анализа поведения пользователей и предотвращения внутренних утечек данных включает следующие программные компоненты:

1. Набор REST веб-сервисов (REST API), который предоставляет методы для отправки статистики и выполнения запросов к уже обработанным данным.
2. Веб-портал, включающий средства управления программной системой и предоставляющий доступ к веб-инструментам визуализации.
3. Обработчик поведенческой статистики, основной функциональностью которого является создание и обновление профилей пользователей.
4. Инструменты визуализации, которые могут быть реализованы как JavaScript приложения, исполняемые в домене Веб-портала, или как внешние программные приложения, обращающиеся к данным через REST API.
5. Поведенческие триггеры. Данные компоненты при наступлении событий отправляют поведенческую информацию для дальнейшей обработки через REST API. Триггеры могут быть реализованы в виде дополнений к существующим программным приложениям или быть развернуты как отдельные программные компоненты.

Поведенческий профиль

Поведенческий профиль является основной сущностью, которой оперирует данная программная система. Операционная система и используемые ппользователем программные при ложения являются источниками данных, из которых в дальнейшем
создаются профили (рис. 1). При создании профиля учитываются только значимые действия. Значимыми действиями могут быть использование справочников, отчетов, проведение бухгалтерских операций, добавление нового документа, удаление документа и др. К данной категории не относятся движения мыши, нажатия клавиш, измение размеров окна и т.д. Профиль пользователя хранится в формате JSON (рис. 3).

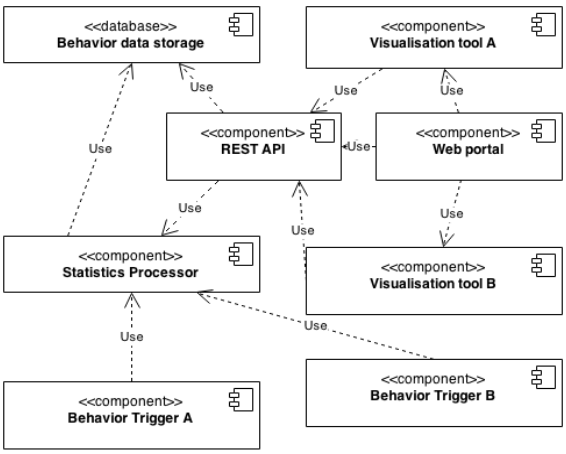

Рис. 1. Компоненты программной системы

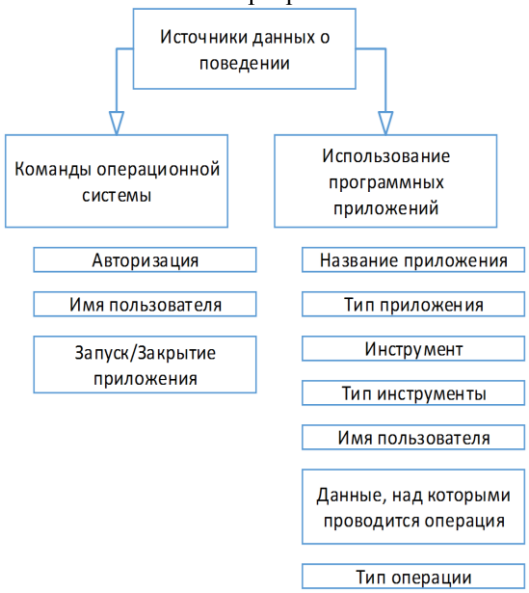

Рис. 2. Источники данных о поведении пользователей

```
{
   "user": "Svetlana",
   "tool_types":
        ["prebuild_report":1032,
         "catalog":5323,
         ..........
         ],
  "tool_names":
         ["daily sales report":542,
         ..................
         ],
  "entities": ["sales": 2093, "employees":3021, "shops":1023
  "operations":"view_col",
  "applications": "1C_Retail",
  "data":
     {.....................    }
}
```

Рис. 3. Пример поведенческого профиля

Над поведенческим профилем выполняются следующие операции:

1. Сложение. Данная операция создает новый поведенческий профиль на основе нескольких имеющихся. Данная операция используется для создания профилей, описывающих более долгосрочные характеристики поведения пользователя, и позволяет отследить процесс эволюции профилей (рис. 4).

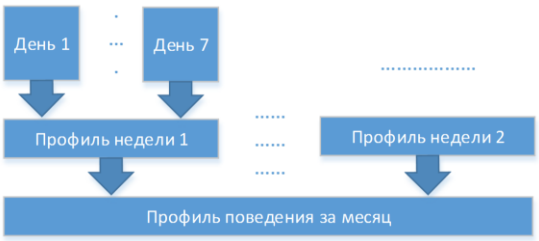

Рис. 4. Сложение поведенческих профилей

2. Нормализация. Необходимость данной операции обусловлена тем, что в профилях хранятся абсолютные значения за различные временные интервалы.

3. Сравнение. Для выполнения операции сравнения нормализованных профилей используется формула расстояния Минковского:

$$d(x_i, x_j) = \sqrt{|x_{i1} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2}$$

Характеристики профилей пользователей

Для профилей также рассчитываются следующие метрики:

1. Волатильность - мера изменчивости поведения пользователя или группы пользователей.

2. Уровень активности - мера, показывающая частоту совершения статистически значимых действий.

3. Операционный тип относит пользователя к одной из следующий групп: читатель, редактор, создатель, неопределенный

Предотвращение внутренних утечек данных

Предотвращение внутренних утечек данных происходит в два этапа:

1. Проверка действия пользователя на соответствие набору статических правил (рис. 5).

2. Проверка степени соответствия краткосрочного профиля пользователя его долгосрочным профилям и долгосрочным профилям группы данного пользователя.

Статические правила делятся на два типа: исключающие и неисключающие. Исключающее правило прерывает процесс проверки и возвращает единственную и однозначную оценку действия пользователя. Примером исключающего правила может быть правило, созданное специально для руководителя организации или системного администратора. По выполнении проверки действию пользователя присваивается итоговая оценка.

```
<rule name="Sales_rule" scope="stat_entry">
    <application>1C</application>
    <userGroup>Operator</userGroup>
    <tool>ProductProviders</tool>
    <operations>
      <read/>
    </operations>
    <time>
        <from> ....</from>
        <to>....</to>
    </time>
    <action>
       <contribute>50</contribute>
    </action>
    <stat>
      <count>84</count>
    </stat>
</rule>
```

Рис. 5. Пример статического правила.

Все статические правила хранятся в виде набора XML-документов.

Заключение

Дальнейшими работами по реализации программной системы является создание классификаторов на основе нечеткой логики и визуализация их работы. Помимо этого, требуется реализовать механизмы авторизации подключаемых инструментов визуализации. Также значительный объем поведенческой статистики требует реализации параллельной и асинхронной обработки для достижения большего быстродействия.

Литература

[1] Hao Wei; Xingyuan Chen; Chao Wang, "User behavior analyses based on network data stream scenario," Communication Technology (ICCT), 2012 IEEE 14th International Conference on , vol., no., pp.1017,1021, 9-11 Nov. 2012.

[2] Giroux, S.; Wachowiak-Smolikova, R.; Wachowiak, M.P., "Keystroke-based authentication by key press intervals as a complementary behavioral biometric," Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on , vol., no., pp.80,85, 11-14 Oct. 2009.

[3] Iglesias, J.A; Ledezma, A; Sanchis, A, "Evolving systems for computer user behavior classification," Evolving and Adaptive Intelligent Systems (EAIS), 2013 IEEE Conference on , vol., no., pp.78,83, 16-19 April 2013.