Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

Научно-исследовательская работа по теме: Классификация существующих методов анализа пользовательской активности

Студент: Пронин Арсений Сергеевич
Группа: ИУ7-72Б
Руководитель: Никульшина Татьяна Александровна

2023 г.

Цель и задачи

Цель: провести обзор существующих методов анализа пользовательской активности, сформулировать критерии для их оценки и провести сравнение рассмотренных методов.

Задачи:

1) рассмотреть существующие решения в области анализа пользовательской активности;
2) классифицировать методы анализа пользовательской активности;
3) выбрать для них критерии оценки и сравнить.

Пользовательская активность

Пользовательская активность это действия совершаемые пользователем при взаимодействии с интерфейсом программы (движение курсора мыши, нажатие клавиш мыши, нажатие клавиш клавиатуры и т.д.), и их характеристики (координаты курсора, частота нажатия, используемые клавиши и т.д.).

Тестирование удобства использования программного обеспечения обычно состоит из двух этапов:

1) сбор данных о пользовательской активности;
2) анализ этих данных.

Математическая модель пользовательской активности ПО

Сессия – последовательность действий пользователя за фиксированный временной промежуток.

Шаблон – предопределенная последовательность событий.

Поддержка шаблона сессией – процент содержания этого шаблона в данной сессии.

Например, пусть имеется сессия $\langle 2, 1, 2, 1, 3, 2, 1, 2, 1, 3\rangle$.
Рассчитаем кол-во вхождений (μ) и поддержку (λ) для следующих шаблонов:

$p1 = \langle 2, 1\rangle$, $\mu = 4$, $\lambda = 0.8$;
$p2 = \langle 2, 1, 2, 1\rangle$, $\mu = 2$, $\lambda = 0.8$;
$p3 = \langle 2, 1, 2, 1, 3\rangle$, $\mu = 2$, $\lambda = 1$;
$p4 = \langle 3, 2, 1, 2, 1\rangle$, $\mu = 1$, $\lambda = 0.5$.

Пример работы алгоритма Apriori

TransID	ItemsPurchased
101	Молоко, Хлеб, Яйца
102	Молоко, Сок
103	Сок, Масло
104	Молоко, Хлеб, Яйца
105	Масло, Яйца

Item	Num
Молоко	1
Хлеб	2
Яйца	3
Сок	4
Масло	5

ID	Items
101	1,2,3
102	1,4
103	4,5
104	1,2,3
105	5,3

Уровень поддержки набора продуктов (support) показывает процент транзакций содержащих набор. Зададим минимальный уровень поддержки 25%.

ItemSet	Support	%
{1}	3	60
{2}	2	40
{3}	3	60
{4}	2	40
{5}	2	40

ItemSet	Support	%
{1,2}	2	40
{1,3}	2	40
{1,4}	1	20
{1,5}	0	0
{2,3}	2	40
{2,4}	0	0
{2,5}	0	0
{3,4}	0	0
{3,5}	1	20
{4,5}	1	20

ItemSet	Support	%
{1,2,3}	2	40

ItemSet	Support	%
∅	0	0

Уровень уверенности (confidence) показывает на сколько вероятно срабатывает полученное правило.

$$\frac{supp(\{1,2,3\})}{supp(\{2\})} = \frac{2}{2} = 100\%$$

Rule	Conf.
Хлеб → {Молоко, Яйца}	100%

Алгоритм GSP

- является модификацией алгоритма AprioriAll;
- учитывает ограничения по времени между соседними транзакциями и id клиента совершившего ее;
- поддержка последовательности - это отношение числа покупателей, в чьих транзакциях присутствует указанная последовательность к общему числу покупателей.

В работе алгоритма можно выделить следующие основные этапы:

1. Генерация кандидатов.
   1.1. Объединение.
   1.2. Упрощение.
2. Подсчет поддержки кандидатов.

Пример работы алгоритма GSP

Генерация кандидатов:

Частые 3-последовательности L_3	Кандидаты 4-последовательности C_4	
	После объединения	После упрощения
$\langle\{1, 2\}\ \{3\}\rangle$	$\langle\{1, 2\}\ \{3, 4\}\rangle$	$\langle\{1, 2\}\ \{3, 4\}\rangle$
$\langle\{1, 2\}\ \{4\}\rangle$	$\langle\{1, 2\}\ \{3\}\ \{5\}\rangle$	
$\langle\{1\}\ \{3, 4\}\rangle$		
$\langle\{1, 3\}\ \{5\}\rangle$		
$\langle\{2\}\ \{3, 4\}\rangle$		
$\langle\{2\}\ \{3\}\ \{5\}\rangle$		

Объединение:

$\langle\{1, 2\}\ \{3\}\rangle + \langle\{2\}\ \{3, 4\}\rangle = \langle\{1, 2\}\ \{3, 4\}\rangle$

$\langle\{1, 2\}\ \{3\}\rangle + \langle\{2\}\ \{3\}\ \{5\}\rangle = \langle\{1, 2\}\ \{3\}\ \{5\}\rangle$

Упрощение:

$\langle\{1\}\ \{3\}\ \{5\}\rangle \notin L_3 \Rightarrow$ удаляем $\langle\{1, 2\}\ \{3\}\ \{5\}\rangle$

Подсчет поддержки кандидатов:

Время транзакции	Объекты
10	1
20	3
26	2

Зададим минимальное и максимальное допустимое время между транзакциями min_gap = 5, max_gap = 15 и размер окна win_size = 0.

Последовательность $\langle\{1\}\ \{3\}\rangle$ поддерживается клиентом, а $\langle\{1\}\ \{2\}\rangle$ нет.

При win_size = 6 одно-элементная последовательность $\langle\{3, 2\}\rangle$ поддерживается клиентом, а $\langle\{1, 3\}\rangle$ нет.

Метод оценки эффективности интерфейса GOMS

- включает в себя модель Keystroke-level Model (KLM)
- KLM выделяет элементарные задачи и длительность каждой из них (рассчитанные на основе усредненных данных лабораторных испытаний). Например, P – указание курсором мыши на объект – 1.1 сек. и B – нажатие или отпускание мыши – 0.1 сек.
- оценка времени на решение задачи сводится к сложению продолжительностей каждой из простейших составляющих. Например, задача, состоящая из классов ⟨P, P, B⟩, потребует для завершения 2.3 сек. (1.1 сек. + 1.1 сек. + 0.1 сек.).

Классификация

- поиск ассоциативных правил (Apriori);
- поиск последовательных шаблонов (GSP);
- сбор и анализ временных характеристик выполнения пользователем действий и промежутков между ними (GOMS);
- вычисление уровней поддержки шаблонов поведения пользователя (Математическая модель пользовательской активности ПО).

Сравнение методов

Метод	Требование к входным данным	Учет времени транз-ий	Сложность алгоритма
Мат. модель пользов. актив. ПО	Множество событий, функция классификации событий, множество сессий, множество последовательных шаблонов	Нет	$O(n \cdot m)$, где n – кол-во шаблонов, m – кол-во сессий
Apriori	Транзакции с набором элементов и минимальный уровень поддержки	Нет	$O(\|D\| \cdot \|I\| \cdot 2^{\|I\|})$, где $\|D\|$ – кол-во транзакций, $\|I\|$ – общее число предметов
GSP	База данных с полями: id последовательности, id и время транзакции, набор элементов и минимальный уровень поддержки	Да	$O(\|I\|^l)$, где $\|I\|$ – общее число предметов, l – длина наибольшей ЧВП
GOMS	Последовательность действий	Нет	$O(n)$, где n – число действий в послед-ти

Заключение

По итогу проделанной работы была достигнута **цель -** проведен обзор существующих методов анализа пользовательской активности, сформулированы критерии для их оценки и проведено сравнение рассмотренных методов.

Также были решены все поставленные **задачи**, а именно:

1) рассмотрены существующие решения в области анализа пользовательской активности;
2) классифицированы методы анализа пользовательской активности;
3) выбраны критерии для их оценки и проведено сравнение.