УДК 519.72+681.3

ОБ АДАПТИВНЫХ МЕТОДАХ ДИНАМИЧЕСКОЙ ИДЕНТИФИКАЦИИ ПОЛЬЗОВАТЕЛЕЙ ПРОГРАММНЫХ ПРОДУКТОВ В РАСПРЕДЕЛЕННЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

Е.И. Кучеренко, В.А. Филатов, Л.Э. Чалая
(Харьковский национальный университет радиоэлектроники)

Предложен адаптивный метод динамической идентификации пользователей распределенных информационных систем на основе положений частотного подхода и теории нечетких множеств. Определены пути практических реализаций теоретических положений.

распределенная информационная система, адаптивный метод, динамическая идентификация, частотный подход, нечеткое множество

Введение. Развитие распределенных информационных систем (РИС) и информационных технологий вызывает необходимость совершенствования процедур идентификации пользователей программных продуктов. Поиск решения задачи идентификации ведется параллельно в таких основных направлениях: тестирующие системы для определения профпригодности (например, для операторов, работающих в режиме on-line); системы дистанционного образования; системы безопасности и защиты информации в распределенных информационных системах и др.

Одним из возможных подходов к решению данной проблемы является использование биометрических методов идентификации [1]. Биометрические методы включают статические и динамические подходы [1]. Для реализации этих подходов обычно требуются дополнительные программно-технические средства, что экономически и организационно не всегда целесообразно в условиях практической реализации.

Перспективным направлением в решении этого вопроса является разработка и внедрение методов динамической идентификации пользователей распределенных информационных систем на основе множества наблюдаемых факторов, характерных для объекта идентификации (пользователя РИС). К этим факторам следует отнести: использование характерных действий при реализации некоторых операций в программных оболочках; временные характеристики и последовательности выполнения операций; иные определяющие факторы. По данным экспертов, дополнительных аппаратных средств для реализации таких методов не требуется, а программные затраты минимальны [1].

В настоящее время эффективных решений задачи динамической идентификации пользователей не предложено, что определяет актуальность данного исследования.

Целью данной работы является повышение достоверности процессов принятия решений по проблеме идентификации пользователей программных продуктов на основе разработки и исследования моделей и адаптивных методов анализа поведенческих характеристик («почерка») пользователей.

Постановка задачи. Пусть задано некоторое метрическое пространство, в котором реализуются поставленные перед пользователем цели $\{C_l\}$, $l \in L$. Согласно [1], учитывая ограничения на множестве программных продуктов $\{S_k\}$, $k \in K$, а также особенности функционирования программ (например, встроенные функции, использование функциональных клавиш, использование мыши и др.) для множества пользователей $U_n \in \{U_n\}, n \in N$, действия при реализации поставленных целей характеризуются повторяемостью (типичностью).

Определение 1. Физические действия (нажатия клавиш и кнопок мыши) для выполнения операции будем называть атомарными, а саму операцию считать некоторой последовательностью атомарных действий.

На множестве используемых программных средств $\{S_k\}, k \in K$, операций $\{d_j\}, j \in J$ и атомарных действий для их выполнения $\{a_i\}, i \in I$ в процессе реализации цели $c_l \in \{C_l\}, l \in L$ требуется уникальным образом идентифицировать пользователя $U_n \in \{U_n\}, n \in N$.

При влиянии на пользователя множества внешних факторов $\{\Phi_m\}, m \in M$ может происходить некоторое нарушение повторяемости применяемых пользователем средств [1], что приводит к неопределенности и нечеткости наблюдаемых факторов.

Необходимо предложить формальные критерии и методы решения поставленной задачи.

Подходы к построению методических средств решения задачи. Операцию, выполняемую пользователем в некоторой программной среде, можно задать следующим образом

$$d_j = < U_n, S_k, c_l, \{a_i\}, \tau >, \qquad (1)$$

где $U_n \in \{U_n\}, n \in N$ – идентификатор пользователя; $\{S_k\}, k \in K$ – программный продукт; $c_l \in \{C_l\}, l \in L$ – цель, стоящая перед пользователем, для реализации которой необходимо выполнения данной операции; $\{a_i\}, i \in I$ – множество атомарных действий, используемых при выполнении данной операции; τ – время выполнения операции.

Пусть для реализации заданной цели в рамках некоторого программного продукта $S_k \in \{S_k\}, k \in K$ требуется выполнение некоторого подмножества операций

$$\{d_{j1}\} \subseteq \{d_j\}, j_1 \in J_1, J_1 \subseteq J. \quad (2)$$

Операции из (2) могут быть реализованы некоторым подмножеством атомарных действий

$$\{a_{i1}\}, i_1 \in I_1, I_1 \subseteq I. \quad (3)$$

В общем случае каждым из пользователей $U_n \in \{U_n\}, n \in N$ подмножество операций (2) реализуется типичным для него способом. Это дает возможность утверждать, что типичность выполнения операций (набор определенных атомарных действий и порядок их использования) – индивидуальная особенность каждого пользователя. Таким образом, мы получаем принципиальную возможность решения поставленной задачи.

Например, для того, чтобы распечатать документ в текстовом редакторе Microsoft Office Word можно использовать следующие действия: а) нажать указателем мыши на иконку «печать» на стандартной панели инструментов; б) войти в меню «Файл», подменю «Печать» и, установив параметры печати, нажать кнопку «ОК»; в) нажать клавиши Ctrl+p, установить параметры печати, нажать кнопку «ОК». Если рассматривать действия пользователя для достижения некоторой цели (получения результата) как некоторую последовательность элементарных (атомарных) действий, то с логической точки зрения такую операцию можно считать неделимой: «распечатка текста». С физической же точки зрения, эта операция будет представлена нажатиями клавиш на клавиатуре или кнопок мыши.

Как было отмечено выше, всякая операция реализуется последовательностью атомарных действий (рис. 1). Для идентификации пользователя важна не только последовательность этих действий, но и временные интервалы их выполнения, а также время между выполнением отдельных атомарных действий.

Существующие методы и алгоритмы обработки данных в разной степени чувствительны к размеру и репрезентативности выборки, к качеству, природе и полноте данных. В связи с этим рассмотрим возможные ситуации, возникающие при идентификации пользователя:

1. Существует значительное число наблюдений выполнения операций, когда

$$\left|\{d_j\}\right| \geq \alpha_1, j \in J, \quad (4)$$

где α_1 – некоторое наперед заданное конечное число, при одновременном выполнении

$$\left|\{U_n\}\right| \geq \alpha_2, n \in N, \quad (5)$$

где α_2 – некоторое наперед заданное конечное число.

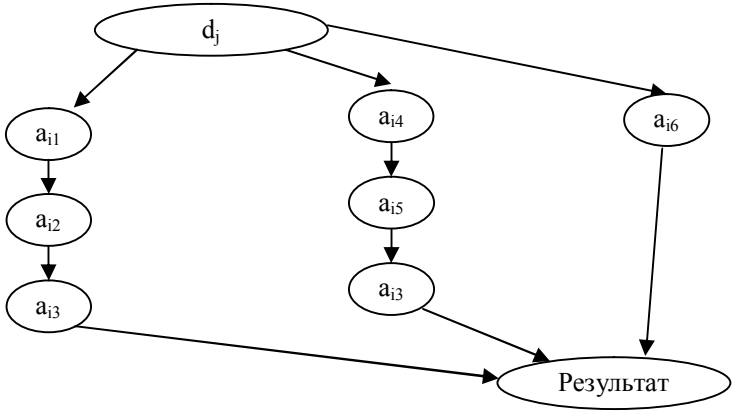

Рис. 1. Иллюстративный пример, характеризующий выполнение операции $\{d_j\}$, $j \in J$, при помощи различных последовательностей атомарных действий

2. Существует незначительное число наблюдений выполнения операций, когда

$$\left|\{d_j\}\right| < \alpha_1, j \in J, \quad (6)$$

при одновременном выполнении (5).

3. Существует число наблюдений выполнения операций, когда

$$\left|\{d_j\}\right| \leq \alpha_2, j \in J, \quad (7)$$

где α_2 составляет единичные наблюдения, при одновременном выполнении (5).

Для случаев 1 и 2 число наблюдений выполнения операций и используемых для их выполнения последовательностей атомарных действий $\{a_i\}, i \in I$ может быть представлено некоторыми случайными величинами с интегральной дискретной функцией распределения [2]. Как показал анализ, при определении и идентификации пользователя целесообразно использовать положения теории вероятности [2].

Для случая 3 положения частотного подхода, используемого в теории вероятности, мало применимы из-за уникальности событий при значении числа наблюдений α, определяемых лингвистической переменной вида «очень малое», «малое». В этом случае целесообразно применение положений теории нечетких множеств [3 – 5], что дает принципиальную возможность получить приемлемые значения показателей достоверности принимаемых решений.

Методы идентификации пользователя на основе наблюдаемых факторов. Анализ ситуаций, возникающих при идентификации стиля работы (почерка) пользователей программных продуктов, выполненный выше, дает возможность построения некоторых стратегий. В связи с этим сформулируем следующее утверждение.

Утверждение 1. Если справедливо (4) и (5), то для каждой последовательности атомарных действий $a_{i1} \in \{a_{i1}\}$, полученной в результате наблюдений выполнения некоторой операции $d_j \in \{d_j\}, j \in J$, формируем идентифицирующие признаки пользователя на основе вычислений математического ожидания $M(x)$ ее реализации.

Действительно, пользователь $U_n \in \{U_n\}, n \in N$ в данном случае может быть определен со значительной степенью уверенности применяемой последовательностью атомарных действий $\{a_{i1}\}, i_1 \in I_1, I_1 \subseteq I$ при выполнении каждой из операций $\{d_{j1}\}, j_1 \in J_1, J_1 \subseteq J$.

Достоверность утверждения 1 очевидна, если учесть, что полученные в практических реализациях результаты при большом α_1 ($\alpha \geq 1000$) имеют значительную сходимость, что подтверждено экспериментом.

Замечание 1. Согласно утверждению 1 идентификация пользователя осуществляется для всех $S_k \in \{S_k\}, \forall S_k, k \in K$.

При малом числе наблюдений α ($\alpha \approx 10$) положения классической теории дают существенную погрешность и незначительную достоверность.

В связи с этим сформируем следующее утверждение.

Утверждение 2. Если справедливо (5) и (6), то для каждой из применяемых последовательностей атомарных действий $a_{i1} \in \{a_{i1}\}, i_1 \in I_1, I_1 \subseteq I$ оценки результатов наблюдений случайных величин должны уточняться на основе распределения Стьюдента [2].

Действительно, при малых α, как следует из статистики малых выборок [3], оценка достоверности событий является достаточно грубой, а сама достоверность малая. Как характерный пример определено, что уточнение результатов на основе распределения Стьюдента при $\alpha = 12$ дает уменьшение ошибки до 10 % [2].

Замечание 2. Результаты динамической идентификации пользователя могут уточняться на основе временных и иных наблюдаемых характеристик.

Для случаев единичных наблюдений (7) реализация положений утверждений 1 и 2 дают недопустимо малое значение достоверности оценок наблюдаемых величин. Это делает затруднительным использование теории вероятности в практической реализации, так как не всегда представляется возможным формирование выборок для пользователя подмножества (3) при реализации (2).

В этом случае перспективным является применение теории нечетких множеств и нечеткой логики [3 – 5].

Для решения поставленной задачи с учетом (7), в работе предлагается следующая стратегия:

1. Для некоторого программного продукта $S_k \in \{S_k\}, k \in K$ определяется подмножество используемых пользователями $\{U_n\}, n \in N$ атомарных действий для реализации операции

$$\forall U_n \in \{U_n\}, \forall d_{j1} \in \{d_{j1}\} \{a_{i1}\} = \text{true}, j_1 \in J_1, J_1 \subseteq J, i_1 \in I_1, I_1 \subseteq I . \quad (8)$$

2. Для каждого пользователя $U_n \in \{U_n\}, n \in N$, определяем

$$U_n \in \{U_n\}, \forall d_{j1} \in \{d_{j1}\} | a_{i1} = \text{true}, j_1 \in J_1, J_1 \subseteq J, i_1 \in I_1, I_1 \subseteq I . \quad (9)$$

Замечание 3. Достоверность идентификации, согласно (8), (9), может оказаться незначительной, что определяется недостаточностью выборки. В связи с этим целесообразно уточнять результаты идентификации на основе временных характеристик выполнения операций.

3. На основе мониторинга выполняемых операций $\{d_{j1}\}, j_1 \in J_1, J_1 \subseteq J$ и использованных для их атомарных действий $\{a_{i1}\}, i_1 \in I_1, I_1 \in I$ определяем время выполнения каждого действия τ^*_{0i1}:

$$\forall \{d_{j1}\}, \{a_{i1}\} | \tau_{i1} = \tau^*_{0i1} . \quad (10)$$

4. Для каждого из зарегистрированных пользователей $\forall U_n \in \{U_n\}, n \in N$ время реакции на выполнение некоторых операций соответствующими средствами может быть представлено в виде конкретных значений времен типа τ_0, а также лингвистических переменных типа «приблизительно», «большое», «малое», «среднее», т.п.

5. Исходя из предметной области, определяем параметры функций принадлежности $\mu(\tau)$ времени τ_{i1} выполнения операций (8) некоторыми средствами (9). Эти функции принадлежности могут быть представлены в виде [3]:

– «малое» значение:

$$\mu_1(\tau) = e^{-k_1 \tau^2}, k_1 > 0 , \quad (11)$$

где k_1 – параметр настройки;

– «очень малое» значение:

$$\mu_2(\tau) = (\mu_1(\tau))^2 , \quad (12)$$

– «среднее» значение:

$$\mu_3(\tau) = e^{-k_2(\tau - a_1)^2} ; \; k_2 > 0, a_1 > 0, \tau > a_1 , \quad (13)$$

где k_2, a_1 – параметры настройки;

– «большое» значение:

$$\mu_4(\tau) = 1 - e^{-k_3(\tau - a_2)^2}\ ;\ k_3 > 0, a_2 > 0, \tau > a_2\,, \qquad (14)$$

где k_3, a_2 – параметры настройки;

– «очень большое» значение:

$$\mu_5(\tau) = (\mu_4(\tau))^2\,. \qquad (15)$$

6. Для пользователя $U_n \in \{U_n\}, n \in N$ представим множество задействованных средств в виде некоторых кортежей: кортеж A_n включает сведения в виде (8); кортеж B_n включает сведения согласно п.4.

7. С учетом (8), (10) – (13) для потенциальных пользователей $\forall U_n \in \{U_n\}, n \in N$, при реализации (9), формируем нечеткое множество задействованных средств в виде

$$\forall \{d_{j1}\}, \{a_{i1}\} \Big| \mu(\tau) = \mu_{i1}(\tau^*_{0i1})\,. \qquad (16)$$

8. Для $U_n \in \{U_n\}, n \in N$ находим индекс нечеткости на основе обобщенного относительного расстояния Хемминга [4]

$$U_n \in \{U_n\} \Big| \nu(\tilde{A}_n) = \frac{2}{|\tilde{A}_n|} \cdot d(\tilde{A}_n; \overline{\overline{\tilde{A}}}_n)\,, \qquad (17)$$

где $\overline{\overline{\tilde{A}}}_n$ – четкое множество, ближайшее к нечеткому. В данном случае $\overline{\overline{\tilde{A}}}_n$ определяется компонентами кортежа A_n.

9. Действия согласно п.п. 6, 7, 8 выполняются для всех зарегистрированных пользователей из $\{U_n\}, n \in N$.

10. Вычисленные значения (17) ранжируются по возрастанию.

Замечание 4. Значение $\nu(\tilde{A}_n)^* = \min\{\nu(\tilde{A}_n)\}, n \in N$ реализует задачу динамической идентификации пользователя, принадлежность которого к некоторому зарегистрированному типу пользователей является максимальной.

11. Действия согласно п.п. 1 – 10 выполняются для всех программных продуктов из $\{S_k\}, k \in K$ и являются основанием для принятия ответственных решений по существу вопроса лицом, принимающим решения (ЛПР).

Замечание 5. ЛПР, кроме приведенных критериев и стратегий, может руководствоваться при принятии решения и другими критериями и соображениями.

В развитие положений замечания следует отметить, что при идентификации пользователя в условиях нечеткого пространства состояний в общем случае следует учитывать также некоторые другие определяющие факторы, например, суммарное время нахождения в одной программной среде, частоту использования тех или иных программных продуктов и другие факторы.

В связи с этим, положения пунктов 3 – 11 расширим на некоторое множество факторов $\{\beta_{\gamma i1}\}, \gamma \in \Gamma$, причем $\beta_{\gamma i1}$ в общем случае включает:

– временные характеристики τ_{i1} использования атомарных действий $\{a_{i1}\}$ реализации соответствующих операций из $\{d_{j1}\}$;

– Za_{i1} – суммарную величину «пробега» манипулятора типа «мышь»;

– Wa_{i1} – характеристики, определяющие особенности обращения, например, к функциональным клавишам и т.п.

Показатели $\beta_{\gamma i1} \in \{\beta_{\gamma i1}\}$ и их значения обычно представлены в нечетком пространстве состояний.

Следствие 1. На основе положений п.п. 3 – 11, процедура динамической идентификации пользователя может рассматриваться как принятие решений на основе нечеткой многофакторной кластеризации в гиперпространстве, размерность которого может быть представлена как $\left|\{\beta_{\gamma i1}\}\right|$.

Практическая реализация. На основе предложенных модели и стратегий была выполнена программная реализация адаптивной системы динамической идентификации пользователей на основе наблюдаемых факторов.

Система идентификации пользователей построена на основе концепции агентных технологий, разработана двухуровневая мультиагентная система идентификации [6, 7], т.е. сама программа не имеет интерфейса, а представлена в виде агента-монитора, и практически реализована как сервис операционной системы (Microsoft Windows XP), для семейства Microsoft Windows 9x – как фоновая программа.

На основе упрощенной модели все атомарные действия при выполнении операции были разделены на три типа: действия путем нажатия клавиш; действия с помощью нажатия правой клавиши «мыши»; действия с помощью нажатия левой клавиши «мыши».

Например, как известно, различные операции в пакете Microsoft Office можно выполнить, применяя средства: сочетания клавиш, манипулятор «мышь». Так, операцию открытия документа можно выполнить, нажав одновременно клавиши Ctrl+O, Ctrl+F12, либо щелкнув «мышью» на значок на верхней панели инструментов, либо, в меню «Файл» вы-

брать команду «Открыть». Пользователь также может назначать удобные для него сочетания клавиш, кроме уже имеющихся (встроенных), в подменю «Настройка».

Эксперимент проходил в условиях учебной лаборатории кафедры искусственного интеллекта. Для исследования были выбраны компьютеры с процессором Celeron с тактовой частотой 1 ГГц, операционной памятью не меньше 256 Мб, свободной памяти на HDD должно быть не менее 1 Гб. В качестве программных продуктов были выбраны широко распространенные программные оболочки, а именно, компоненты программного пакета Microsoft Office XP: Microsoft Access, Microsoft Excel, Microsoft Word. В ходе экспериментальной проверки были задействованы следующие группы пользователей: инженеры-программисты, студенты младших курсов специальностей направления «Компьютерные науки», студенты старших курсов специальностей того же направления.

Тестовая группа состояла из тридцати человек, работала в течение 1 месяца по 3 часа в день. Этот период был разбит на два этапа: обучение системы – 3 недели, и тестирование – 1,5 недели. В среднем каждым пользователем было выполнено порядка 10000 атомарных действий.

В результате работы программы формируется XML-файл специального вида, содержащий подробную информацию о действиях пользователей, и который впоследствии обрабатывается в соответствии с предложенными в данной работе стратегиями.

Результаты эксперимента показали, что для всех задействованных групп пользователей получены хорошие приближения. Так, точность распознавания пользователей, выраженная в процентном отношении, составила около 83,4%. Ошибка, возникшая в процессе тестирования, обусловливается неустойчивостью «почерка» некоторых пользователей, а также использованием упрощенной модели (если операции и атомарные действия разбить на большее количество типов, то точность распознавания значительно повысится).

Перспективным направлением исследования является расширение множества наблюдаемых факторов, а также учет влияния внешних факторов, включая психофизические (например, стрессовые ситуации) на достоверность теоретических посылок.

Выводы. 1. На основе анализа существующих решений сформулирована задача построения и формализации методов идентификации пользователя программных продуктов на основе наблюдаемых факторов.

2. Впервые предложен формализованный адаптивный метод динамической идентификации пользователя на основе вероятностных подходов и нечеткой логики, что позволяет расширить возможности иденти-

фикации пользователя РИС, обладающего различными возможностями в использовании и реализации программных продуктов.

3. Полученные теоретические результаты являются основой построения алгоритмического и программного обеспечения для решения прикладных задач.

4. Практическая значимость и достоверность теоретических результатов подтверждена при тестировании групп пользователей, обладающих различной квалификацией при работе с программно-техни-ческими средствами и различной реакцией на возмущения внешней среды.

5. Сферой эффективного применения полученных результатов является формирование профиля и идентификация пользователя программных продуктов, входящих в состав тренажерных комплексов, испытательных стендов, систем контроля и защиты информации, дистанционного образования и других сфер применения распределенных информационных систем.

6. Перспективными направлениями дальнейших исследований является расширение множества наблюдаемых факторов, которые учитываются при идентификации пользователей, включая нечеткое пространство состояний наблюдаемых факторов.

ЛИТЕРАТУРА

1. *Иванов А.И. Биометрическая идентификация личности по динамике подсознательных движений (монография) – Пенза: ПГУ, 2000. – 188 с.*
2. *Сигорский В.П. Математический аппарат инженера. – К.: Техніка, 1975. – 768 с.*
3. *Кофман А. Введение в теорию нечетких множеств: Пер. с франц. – М.: Радио и связь, 1982. – 432 с.*
4. *Нечеткие множества и теория возможностей. Последние достижения: Пер. с англ. / Под ред. Р.Р. Ягера. – М.: Радио и связь, 1986. – 408 с.*
5. *Tsoukalas L.H., Uhrig R.E. Fuzzy and Neural Approaches in Engineering. – New York: John Wiley&Sons.Inc, 1997. – 587 p.*
6. *Кучук Г.А. Формалізація предметної області багатовимірних баз даних // Системи обробки інформації. – Х. : ХФВ: «Транспорт України», 2001. – Вип. 1(11). – С. 110 - 114.*
7. *Пономаренко Л.А., Филатов В.А. Динамическое администрирование баз данных с использованием агентных технологий // Научный и произв.-практич. сборник "Труды Одесского политехнического университета". – Одесса: ОНПУ. – 2001. – Вып. 4 (16). – С. 95 – 97.*

Поступила 20.02.2005

Рецензент: доктор технических наук, профессор Е.В. Бодянский,
Харьковский национальный университет радиоэлектроники.