УДК 004.75:[004.65]

Л.Э. Чалая

Харьковский национальный университет радиоэлектроники, Харьков

МЕТОД ИДЕНТИФИКАЦИИ ПОЛЬЗОВАТЕЛЕЙ ИНФОРМАЦИОННЫХ СИСТЕМ НА ОСНОВЕ МНОГОСВЯЗНОГО ПРЕДСТАВЛЕНИЯ КЛАВИАТУРНОГО ПОЧЕРКА

В статье рассматривается задача идентификации пользователей компьютерных систем на основе многосвязного представления особенностей их клавиатурного почерка. Предложена процедура классификации, позволяющая осуществлять биометрическую идентификацию в реальном масштабе времени. Приведены результаты экспериментальных исследований работы биометрической системы.

биометрическая идентификация, информативные параметры, полиграфы, клавиатурный почерк

Введение

Важной предпосылкой повышения эффективности динамической идентификации пользователей компьютерных систем является выбор методов, позволяющих получить адекватное и компактное описание задачи такой идентификации. Для распределенных информационных систем (РИС) получили распространение методы, основанные на анализе особенностей клавиатурного почерка пользователей.

Разработка процедур выбора и последующего анализа информативных параметров биометрической системы динамической идентификации пользователей РИС является нетривиальной задачей. Рассмотрим биометрические параметры, которые могут принципиально использоваться при решении задачи идентификации пользователей по клавиатурному почерку.

Определение. Клавиатурным почерком пользователя будем называть совокупность индивидуальных характеристик, определяющих особенность его работы в режиме ввода текста с клавиатуры (например, время удержания клавиши, времена между нажатиями клавиш, частота использования функциональных клавиш и т.д.).

В соответствии с этим определением при идентификации индивидуального клавиатурного почерка пользователя РИС его первичный образ формируется на основе индивидуальных параметров–характеристик: времени удержания каждой клавиши (ВУК), времени между нажатиями клавиш (ВМК), скорости набора, ритмичности и т.д.

При выборе параметров для анализа клавиатурного почерка можно использовать стратегии методов сжатия данных, а именно, стратегии статического словаря, состоящего из последовательностей букв длины n, где $2 \le n \le 5$. В этом случае при анализе результатов "свободного набора" учитываются полиграфы (биграфы, триграфы, тетраграфы и т.д.), то есть комбинации смежных символов, встречающиеся как в контрольной фразе, которая использовалась для формирования эталонной матрицы, так и идентичные буквосочетания при «свободном наборе», и вычисляются временные зависимости между нажатиями нескольких клавиш в одинаковых сочетаниях.

Предлагаемая в настоящей статье схема аутентификации пользователей РИС предполагает формализацию процедуры анализа их клавиатурного почерка и последующего использования биометрических образов в комбинированной системе.

Постановка задачи

Априорной информацией при оценке информативности отдельных ВМКn для полиграфов n-го порядка, используемых в качестве биометрических параметров пользователей РИС, может служить статистика о частоте формирования полиграфов в русском и английском языках [1]. Анализ частоты формирования диграфов для украинского языка представляется особенно важным, так как документооборот на государственных предприятиях в настоящее время ведется на украинском языке. Перечень диграфов, наиболее часто встречающихся в технической и служебной документации, используемой в РИС различного функционального назначения, представлен (в порядке убывания вероятности их появления) в табл. 1.

Использование такой статистики позволяет привести входные данные в виде времени между нажатиями клавишей для соответствующих диграфов (ВМК2) к некоторому структурированному виду. Такое представление данных необходимо для уточнения особенностей клавиатурного почерка, лежащего в основе формирования биометрических эталонов и реализации процедуры динамической идентификации пользователей в процессе функцио-

нирования РИС. Помимо этого, на основе анализа особенностей динамики работы на клавиатуре можно судить о стабильности и наличии клавиатурного почерка у пользователей как такового.

Таблица 1

Перечень наиболее используемых диграфов в английском, русском и украинском языках

Диграфы английского языка	Диграфы русского языка	Диграфы украинского языка
th, er, in, ou, an, en, ea, or, ll, is, on, ar, st, gh, ed, ee, om, oo, ow, ss, ur, ld, at, sh, id, sa, ic, tr, al, il, as, ir, ec, ul, ly, et, ai, ch, ot, it, av, im, ol, to, qu	то, ст, ов, ен, по, аз, ер, ол, ор, он, ел, ет, ам, от, ом, ас, ан, ин, ск, на, за, ар, ик, пр, ев, ив, ит, ил, ед, ем, ть, ал, ат, ав, ся, ес, об, од, ос, ис, ог, им, ег, ич, сь	ст, ен, на, ан, та, ни, но, ов, ор, по, ро, ра, ти, за, ко, нн, ре, то, пр, ер, ви, ат, ис, ри, ва, ня, ос, ли, да, те, ть, ці, ль, ом, ал, аб, ка, об, ол, ні, ит, ля, ів, од, ме, их, ик

Известные методы анализа клавиатурного почерка основываются на предположении, что клавиатурный почерк представляется в виде усредненных значений событий клавиатуры. Но, как показывают эксперименты, значения ВУК, ВМК2 в разных комбинациях являются различными, так как времена удержаний и пауз между удержаниями клавиш зависят от предыдущих событий клавиатуры. Если предположить, что клавиатура состоит всего из одной клавиши, то в соответствии с приведенным выше определением клавиатурного почерка естественно предположить, что все удержания и паузы между удержаниями для единственной клавиши будут приблизительно одинаковы между собой. В этом случае значения удержаний и пауз между удержаниями будут устойчивыми только в полных последовательностях кода. Таким образом, можно говорить о том, что информативностью обладают не только значения времен удержаний и пауз между удержаниями клавиш, но и устойчивые последовательности сочетаний значений этих времен.

Отметим, что на этапе предварительного анализа входных данных необходимо исключить из рассмотрения параметры, обладающие малой информативностью и не отражающие особенностей клавиатурного почерка пользователей РИС. Следует выделить два основных подхода к построению информативной системы параметров [2].

Первый подход основан на эвристическом определении малого числа параметров большой информативности. Используемые при этом методы базируются на эвристике и эмпирике, т.е. выбор признаков определяется интуицией и опытом разработчика.

Второй подход заключается в том, что из большого числа исходных параметров по некоторому критерию их информативности выбирается как можно меньшее число наиболее полезных для распознавания параметров.

Предположим, что значения событий клавиши распределены по нормальному закону и характеризуются несколькими центрами распределения (классами).

Предлагаемый ниже метод многосвязного представления особенностей динамики работы на клавиатуре позволяет в полной мере выявить и представить характерные особенности клавиатурного почерка на основе устойчивых последовательностей событий клавиатуры. Для выявления таких структур необходима определенная статистика, состоящая из значений времен событий клавиатуры и связей между событиями клавиатуры [3].

Таким образом, целесообразным представляется разработка автоматического классификатора, составляющего ядро системы идентификации и мониторинга действий пользователей информационных систем, работающего в реальном времени и учитывающего как значения ВУК, ВМК2 для наиболее часто используемых диграфов, так и устойчивые связи между этими характеристиками.

Решение задачи

В общем случае входные данные можно представить как упорядоченное множество событий клавиатуры M. Положение события во множестве определяется порядковым номером появления самого события. Тогда задача многосвязного представления особенностей динамики работы на клавиатуре заключается в разделении исходного множества M на подмножества $A_i, i \in [1, N]$ последовательностей событий клавиатуры, согласно установленным критериям. Причем справедливы следующие выражения:

1. $\forall A_i \quad A_i \subset M,\ i \in [1, N]$.
2. $\exists A_i, A_j\ A_i \cap A_j \neq \varnothing,\ i, j \in [1, N], i \neq j$.
3. $\forall A_i, A_j \quad A_i \neq A_j,\ i, j \in [1, N],\ i \neq j$.

В результате операции разделения множества M образуются два подмножества – A и G, причем $M = A \bigcup G$. Подмножество A является объединением всех подмножеств последовательностей событий клавиатуры

$$A = \bigcup_{i=1}^{N} A_i .$$

Подмножество G содержит все события, которые не удовлетворили установленным критериям, следовательно $G = M \setminus A$. Подмножество G может быть пустым, тогда $M = A$. В этом случае все элементы исходного множества соотнесены определенным последовательностям событий клавиатуры.

Задача разделения исходного множества M на подмножества A и G может быть решена в рамках

подхода автоматической классификации. В результате исходная выборка событий клавиатуры будет разделена на классы для каждой клавиши в отдельности, а связи между событиями клавиатуры исходной выборки будут определять взаимосвязи между полученными классами. На рис.1 схематично показано многосвязное представление особенностей динамики работы на клавиатуре.

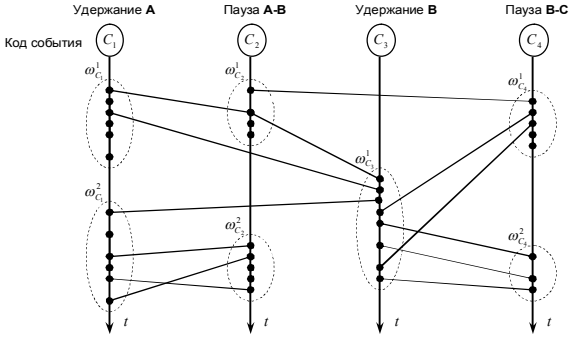

Рис. 1. Многосвязное представление клавиатурного почерка пользователя РИС

Всем событиям клавиатуры исходной выборки в соответствие ставится код события C_i. В зависимости от кода C_i значения событий наносятся на одну из временных осей с кодом C_i.

Связи, определяющие последовательность появления событий клавиатуры при наборе, обозначаются схематично в виде отрезков соединяющих соответствующие значения событий исходной последовательности. Согласно установленным критериям все события с кодом C_i образуют классы $\omega^j_{C_i}$. Элементы класса $\omega^j_{C_i}$ распределены по закону, который приближенно можно считать гауссовым. Полученные классы $\omega^j_{C_i}$ характеризуются центром распределения и среднеквадратичным отклонением. Связь между классами образуется на основе существующих связей между элементами классов и характеризуется мощностью связи L или количеством связей элементов.

Рассмотрим последовательный метод классификации пользователей, использующий в качестве биометрического эталона многосвязное представление особенностей динамики работы на клавиатуре. Предложенный метод классификации основывается на анализе устойчивых последовательностей событий клавиатуры, которые образуются во время работы пользователя на компьютере. В качестве числовой оценки таких последовательностей событий предлагается использовать вес последовательности Q, который зависит от множества значений событий клавиатуры $T = \{t_1, t_2, \dots t_N\}$ и множества кодов событий клавиатуры $C = \{C_{k_1}, C_{k_2}, \dots, C_{k_N}\}$ в исходной последовательности, где N – длина исходной последовательности (количество событий клавиатуры в последовательности), а $k_i, i \in [1, N]$ - определяет индекс кода события. Тогда вес устойчивой последовательности определяется следующим образом

$$Q(C,T) = p_{C_{k_1}}(t_1) + \sum_{i=2}^{N}\left(1 - e^{-\alpha L_i}\right) p_{C_{k_i}}(t_i),$$

где i – номер события клавиатуры в последовательности событий; L_i – мощность связи или количество связей между i-м и $(i-1)$-м событиями клавиатуры; α – коэффициент, определяющий влияние мощности связи на вес Q; $p_{C_{k_i}}(t_i)$ – смесь плотностей вероятности классов событий клавиши с кодом C_{k_i} для события t_i.

Смесь плотностей вероятности определяется как

$$p_{C_{k_i}}(t) = \sum_{j=1}^{M} P\left(\omega^j_{C_{k_i}}\right) p\left(t / \omega^j_{C_{k_i}}\right),$$

где $\omega^j_{C_{k_i}}$ – класс значений времен событий с кодом C_{k_1}; $P\left(\omega^j_{C_{k_i}}\right)$ – априорная вероятность класса $\omega^j_{C_{k_i}}$; $p\left(t / \omega^j_{C_{k_i}}\right)$ – условная плотность вероятности для класса $\omega^j_{C_{k_i}}$.

Вес устойчивой последовательности событий клавиатуры с длиной $N = 1$ равен свободному члену, т.е.

$$Q(C,T) = p_{C_{k_1}}(t_1).$$

Априорная вероятность $P\left(\omega^j_{C_i}\right)$ класса $\omega^j_{C_i}$ событий клавиатуры с кодом C_i определяется как отношение количества событий класса $\omega^j_{C_i}$ к общему количеству событий с кодом C_i в исходной статистике. Условная плотность вероятности для класса $\omega^j_{C_i}$ задается следующим образом:

$$p\left(t / \omega^j_{C_i}\right) = \frac{1}{\sqrt{2\pi D^j_{C_i}}} \exp\left(-\frac{\left(t - m^j_{C_i}\right)^2}{2D^j_{C_i}}\right),$$

где $m^j_{C_i}$, $D^j_{C_i}$ – математическое ожидание и дисперсия класса $\omega^j_{C_i}$ соответственно.

Так как вес последовательности является дискретной величиной, то решающее правило классификации, лежащее в основе процедуры идентификации пользователей по клавиатурному почерку, можно построить на текущем анализе функции $\upsilon(\tau)$:

$$\upsilon_{i+1} = \upsilon_i + \left(bQ(T_i, C_i) - a\upsilon_i\right)\Delta\tau,$$

где a , b – некоторые коэффициенты, которые определяются экспериментально.

Значительные увеличения υ_{i+1} будут определяться появлением устойчивых последовательностей с большим весом и большей длиной. Последовательности единичной длины не приводят к существенному увеличению значений υ_{i+1}, а последовательности с малым весом, характерные для идентификации «чужого», будут резко понижать значение скорости изменения υ_{i+1}. Таким образом, величина изменения скорости υ_{i+1} может характеризовать особенности динамики работы пользователя на клавиатуре.

Для работы классификатора необходимо определить величину среднего значения I изменения величины υ_{i+1} для «своего» и задать верхний υ_1 и нижний υ_2 пороги, которые определяют величину накопления ошибки идентификации (рис. 2).

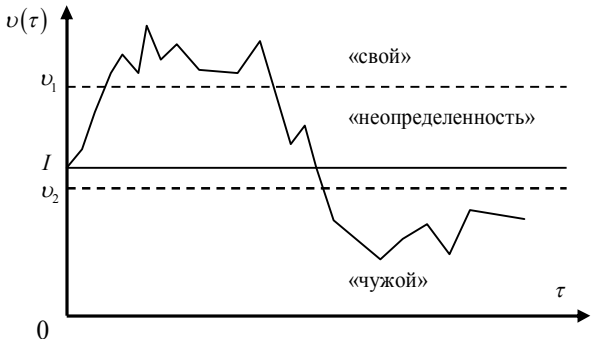

Рис. 2. Классификация пользователей
по последовательному анализу клавиатурного почерка

Тогда решающее правило для последовательного классификатора на основе многосвязного представления клавиатурного почерка можно записать следующим образом:

$$\begin{cases} \upsilon_1 \le \upsilon_{i+1} \rightarrow \text{"свой"}; \\ \upsilon_2 < \upsilon_{i+1} < \upsilon_1 \rightarrow \text{произвести следующее наблюдение}; \\ \upsilon_{i+1} < \upsilon_2 \rightarrow \text{"чужой"}. \end{cases}$$

Среднее значение I величины υ_{i+1}, а также верхний υ_1 и υ_2 нижний пороги классификатора определяются на этапе обучения.

В отличие от существующих последовательных методов классификации (последовательный критерий Вальда, последовательная решающая процедура Байеса и др.) предложенный метод отличается непрерывностью процесса классификации, что позволяет реализовать непрерывную процедуру биометрической идентификации пользователей РИС по клавиатурному почерку.

Экспериментальные результаты

Тестирование работы группы пользователей информационно-аналитической системы ИАС «Университет», позволило получить следующие результаты:

– в ходе эксперимента были собраны и проанализированы временные характеристики клавиатурного почерка 14 пользователей в процессе заполнения ими базы данных ИАС «Университет». По скорости набора текста пользователи были разделены на три кластера. Для обучения системы каждому из пользователей необходимо было набрать предложенный текст (один и тот же), длиной 1000 символов не менее 30 раз. На основании полученных данных для пользователей были построены биометрические образы, основанные на математических ожиданиях ВУК и ВМН. Последующее тестирование проводилось в процессе заполнения пользователями базы данных ИАС «Университет». Правильная идентификация с использованием предложенного в настоящей статье метода в тестовой выборке составила 89,27%;

– для подтверждения эффективности предложенного метода был также проведен эксперимент, в котором было запланировано обучение и тестирование системы идентификации пользователей по клавиатурному почерку без многосвязного представления их биометрических характеристик. Правильная идентификация с использованием метода потенциальных функций и квадратичного индекса нечеткости в данной тестовой выборке составила лишь 83,22%, что подтверждает целесообразность учета связей между биометрическими характеристиками пользователей при создании систем идентификации и мониторинга действий пользователей РИС [4];

– применение варианта биометрической идентификации пользователей ИАС «Университет», основанного на многосвязном представлении характеристик ВМК2, подтвердило снижение вероятности появления ошибок 1-го и 2-го рода соответственно на 0.08 и 0.055 (по сравнению с результатами классической аутентификации).

Список литературы

1. Герасименко В. А., Малюк А.А. Основы защиты информации. – М.,1994. – 540 с.

2. Дубровин В.И., Субботин С.А. Алгоритм классификации с оценкой значимости признаков // Радіоелектроніка. Інформатика. Управління. – 2001. – № 2. – С. 145-150.

3. Брюхомицкий Ю.А., Зверев Е.М., Казарин М.Н. Идентификация пользователя по клавиатурному почерку на основе сменных виртуальных клавиатур // Электронный журнал «Информационное противодействие угрозам терроризма». – 2005. – № 4. – С. 141-146.

4. Чалая Л.Э. Модель идентификации пользователей по клавиатурному почерку // Искусственный интеллект. – 2004. – № 4. – С. 811–817.

Поступила в редколлегию 11.10.2007

Рецензент: д-р техн. наук, проф. О.Г. Удовенко, Харьковский национальный университет радиоэлектроники, Харьков.