В.Г. ШЕРСТЮК

Херсонский национальный технический университет, г. Херсон

Метод динамической оценки подобия двух потоков событий

УДК 004.986

В работе рассмотрена задача оценки подобия последовательностей событий, формируемых в результате мониторинга состояния сложной динамической системы. Предложена формальная модель событийной системы, представлен метод динамической оценки подобия двух потоков событий, основанный на принципе максимально возможного совмещения, показана его адаптивность к условиям неполной и неточной исходной информации, дана оценка эффективности. Метод может быть использован в системах реального времени.

Взаимодействие множества управляемых динамических объектов на ограниченном пространстве образует сложную динамическую систему (СДС) [1]. В СДС зачастую возникает необходимость решения задач диагностики и предсказания нежелательных (критических или аварийных) ситуаций.

Состояние СДС и поведение объектов в ней, как правило, оценивается в процессе непрерывных наблюдений (мониторинга), при этом исходная информация о состоянии СДС представляет собой упорядоченную во времени последовательность событий.

Каждое событие интерпретируется как составной объект, состоящий из множества количественных или качественных оценок параметров, получаемых путем прямых либо косвенных измерений. Присутствие или отсутствие во временной последовательности событий определенного класса может быть основанием для выводов о возможном переходе СДС в то или иное состояние, о вероятных будущих событиях или поведении объектов.

Ограниченная точность средств измерений в системах реального времени, наличие нескольких независимых каналов наблюдения, присутствие шумов и искажений приводят к неполноте, неточности и противоречивости информации о событиях. Как следствие, в наблюдаемом потоке событий могут присутствовать шумовые события, пропуски событий, искажения параметров и т.д.

Выявление причинно-следственных зависимостей между событиями и комбинациями событий в СДС является нетривиальной задачей, требующей участия в ее решении человека-оператора или эксперта.

Снизить зависимость от «человеческого фактора» можно путем автоматизации процессов диагностики и предсказания ситуаций в СДС с использованием интеллектуальных систем (ИС).

Применение ИС, основанных на правилах либо на моделях, по причине невозможности априорного построения адекватных систем правил и моделей, а также ввиду необходимости верификации знаний, неосуществимой во многих открытых предметных областях, для решения поставленной задачи практически невозможно [2].

Наиболее подходящим инструментом могли бы стать ИС, основанные на прецедентах, действующие на основе принципов: «ситуациям свойственно повторяться» и «в подобных ситуациях могут быть приняты подобные решения» [3].

Однако, использовать существующие модели прецедентных ИС для решения задач диагностики и предсказания нежелательных ситуаций невозможно из-за их чрезмерной статичности: независимые друг от друга проблемные ситуации в них никогда не пересекаются во времени – в каждый момент рассматривается только один прецедент, не допускающий развития ситуации и являющийся статическим «снимком» значений параметров.

В то же время, реализация динамических прецедентных ИС требует соответствующего теоретического обоснования [4].

Основой для принятия решений по прецедентам в прецедентных ИС реального времени является выявление подобия для последовательностей (потоков) событий, – наблюдаемых и эталонных, хранимых в рабочей памяти.

Использование событийных моделей, описание динамики СДС в прецедентах с помощью потоков событий требуют разработки мето-

да оценки подобия последовательностей, способного учитывать неполноту и неточность информации, с достаточно низкой для систем реального времени вычислительной сложностью.

Подходы к сопоставлению последовательностей объектов

Существует значительное число работ, посвященных исследованию декларативных и процедурных методов оценки подобия различных объектов; систематический их обзор дан в [5].

Наибольшее распространение в прецедентных ИС получили абсолютные, относительные и метрические оценки подобия, восходящие к известной из [6] теории подобия Тверского и основанные на исследовании совпадающих и различающихся свойств пары объектов. Каждый из объектов представляется как множество числовых параметров, между которыми измеряется расстояние (Евклидово, Манхеттенское, Римановское и т.д.).

Метрические оценки основываются на использовании отношений совершенного или полного порядка для заданного множества и представляются в виде нормализованного вещественного числа из диапазона [0..1], или элементом решетки, построенной на основе отношения порядка.

Декларативные методы метрической оценки подобия имеют значительную вычислительную сложность, вследствие чего не могут использоваться в системах реального времени [7].

Процедурные методы используют искусственные концепты, такие как векторы коэффициентов значимости для совпадающих и различающихся параметров, коэффициенты выпуклости множеств параметров, граничные срезы отношения порядка и т.д., которые, во-первых, являются статическими, а во-вторых, выбираются экспертами в процессе разработки ИС, что существенно ограничивает их применимость [8].

Все рассмотренные декларативные и процедурные методы оценки подобия основаны на принципе попарного сопоставления объектов. Между тем, в динамических ИС требуется оценивать подобие не пар, а последовательностей объектов.

В рассматриваемых задачах необходимо оценивать подобие не отдельных событий, а их последовательностей, поэтому известные методы оценки подобия объектов непосредственно использоваться не могут. В то же время, вопросы оценки подобия последовательно-

стей объектов на сегодняшний день исследованы недостаточно. Так, в [9] предложен метод сравнения длинных последовательностей, основанный на подсчете расстояния Левенштайна, практическая реализация которого затруднительна ввиду его значительной вычислительной сложности.

В [10] для сравнения последовательностей предложены нейронные сети и генетические алгоритмы, которые в ИС реального времени не могут быть использованы в силу их нелинейности.

В [11] предложен подход к сравнению последовательностей объектов, неадаптивный к условиям неточной и противоречивой информации. Идея динамической оценки подобия последовательностей предложена в [12], однако практического воплощения она не получила.

Подходящим для реализации в условиях неполной и неточной информации является метод динамического подсчета вхождений [13], разработанный для решения специфической задачи обнаружения вторжений в компьютерные сети.

В его основе лежит статистическая модель подсчета вхождений событий, отдающая приоритет слабо схожим редко встречающимся последовательностям впротивовес сильно схожим часто встречающимся последовательностям, что препятствует использованию данного метода в других предметных областях..

Таким образом, задача динамической оценки подобия последовательностей событий представляет собой недостаточно исследованную область, решение ее в условиях неполноты и неточности наблюдений в реальном времени известными методами получить невозможно.

Цель работы состоит в синтезе метода динамической оценки подобия двух потоков событий в условиях неполной и неточной информации, пригодного для практической реализации в прецедентных системах реального времени для решения задач диагностики и предсказания ситуаций в СДС.

За основу можно принять [13]. В то же время, необходимо освободиться от статистической модели при оценке подобия отдельных событий, входящих в последовательность. Это возможно реализовать, используя таксономическую иерархию событий [4].

Для обеспечения требуемой эффективности длинные потоки событий также необходимо сегментировать, разбивая на последовательности небольшой длины. Кроме того, необходимо усовершенствовать схему подсчета затрат на совмещение сегментов, обеспечив ей более низкую вычислительную сложность, и, в отличие от [13], обеспечить учет временных взаимоотношений между событиями, для чего задать явную временную шкалу, к которой и привязать события потоков данных.

Структура событийной модели

Формализуем базовые понятия события и потока событий, взяв за основу [13] и вводя в модель параметр времени.

Событийной моделью E назовем тройку:

$$E = \langle \upsilon, r, \mathcal{Z} \rangle, \quad (1)$$

где υ – множество переменных модели;

r – множество ограничений;

$\mathcal{Z}$ – сигнатура.

Иерархия событий $\Im_i$ представляет собой тройку:

$$\Im_i = \langle \perp_i, I_i, \prec_i \rangle, \quad (2)$$

где I_i – множество элементов иерархии, соответствующей определенному отношению $\mathfrak{w}_i$ между событиями;

$\prec_i$ – отношение частичного порядка, заданное над I_i;

$\perp_i$ – наименьший элемент последовательности $\prec_i$;

Сигнатурой событийной модели $\mathcal{Z}$ назовем кортеж вида:

$$\mathcal{Z} = \langle X, \{\Im_1 ... \Im_m\}, T, <_T \rangle, \quad (3)$$

где X – множество параметров событий;

$\{\Im_1 ... \Im_m\}$ – множество иерархий событий $\Im_i$;

T – множество значений времени;

$<_T$ – отношение полного порядка для T.

Пусть I_1 соответствует множеству классов $C = \{c_1, c_2, ... c_n\}$, где c_i – класс события. Тогда отношение частичного порядка $\prec_1$ является отношением информационной упорядоченности, т.е. $c_1 \prec_1 c_2$ означает, что класс c_1 несет в себе меньше информации, чем c_2.

Таким образом, класс c_1 является *абстракцией* класса c_2, класс c_2 – *конкретизацией* класса c_1, а отношение $\prec_1$ задает на C таксономическую иерархию.

Соответственно, $\perp_1$ представляет собой *наиболее абстрактный* (содержащий минимум информации) *класс* таксономической иерархии $\langle \perp_1, C, \prec_1 \rangle$.

В рассматриваемой модели элемент $\perp_1$ имеет семантику «любое событие», т. е. $\forall c \in E.\mathcal{Z}.I_1 \quad \perp_1 \prec c$.

Пусть I_2 соответствует множеству $\Sigma = \{z_1, z_2, ...z_n\}$, где z_j – некоторая композиция событий, $z_j = \psi_k \uplus \psi_l$. Тогда отношение частичного порядка $\prec_2$ является по определению отношением вложения, т.е. $z_1 \prec_2 z_2$ означает, что композиция z_2 включает в себя композицию z_1 (и наоборот, композиция z_1 вложена в z_2).

Таким образом, композиция z_2 является *составной*, z_1 – *элементом* композиции z_2, а $\perp_2$ в композиционной иерархии $\langle \perp_2, \Sigma, \prec_2 \rangle$ имеет семантику «любое событие», т. е. $\forall z \in E.\mathcal{Z}.I_2 \quad \perp_2 \prec z$. В дальнейшем изложении примем $\perp_1 = \perp_2 = \perp$.

Отношение $<_T$ задает на множестве значений времени T вполне упорядоченную временную шкалу $\langle t_0, T, <_T \rangle$, для которой имеется начальный момент времени $t_0 \in T$.

Формализация событий

Событием ψ в модели E назовем структуру вида:

$$\psi ::= \langle Y, c, t, \mathfrak{X} \rangle, \tag{4}$$

где Y – метка события, $Y \in E.\upsilon$;

c – класс события, $c \in E.\mathcal{Z}.I_1$;

t – момент наблюдения события ψ, $t \in E.\mathcal{Z}.T$;

$\mathfrak{X}$ – множество слотов мощности n, $\mathfrak{X} = \{\mathfrak{x}_1, \mathfrak{x}_2, ...\mathfrak{x}_n\}$.

С помощью множества слотов $\mathfrak{X}$ представляются параметры события. *Множеством параметров события* ψ будем называть кортеж вида $X = \{x_1, x_2, ..., x_n\}$, где x_i – параметр.

Область возможных значений параметра x_i обозначим как D_{x_i}.

Слотом $\mathfrak{x}_i$ будем называть множество значений i-го поименованного параметра события x_i в определенные моменты времени $t_j \in T$, имеющее структуру:

$$\mathfrak{x}_i = \left\langle \hat{x}_i \doteq \left\{ \mathcal{X}_i^{t_j}, \mathcal{X}_i^{t_{j+1}}, ... \mathcal{X}_i^{t_{j+m}} \right\} \right\rangle, \quad (5)$$

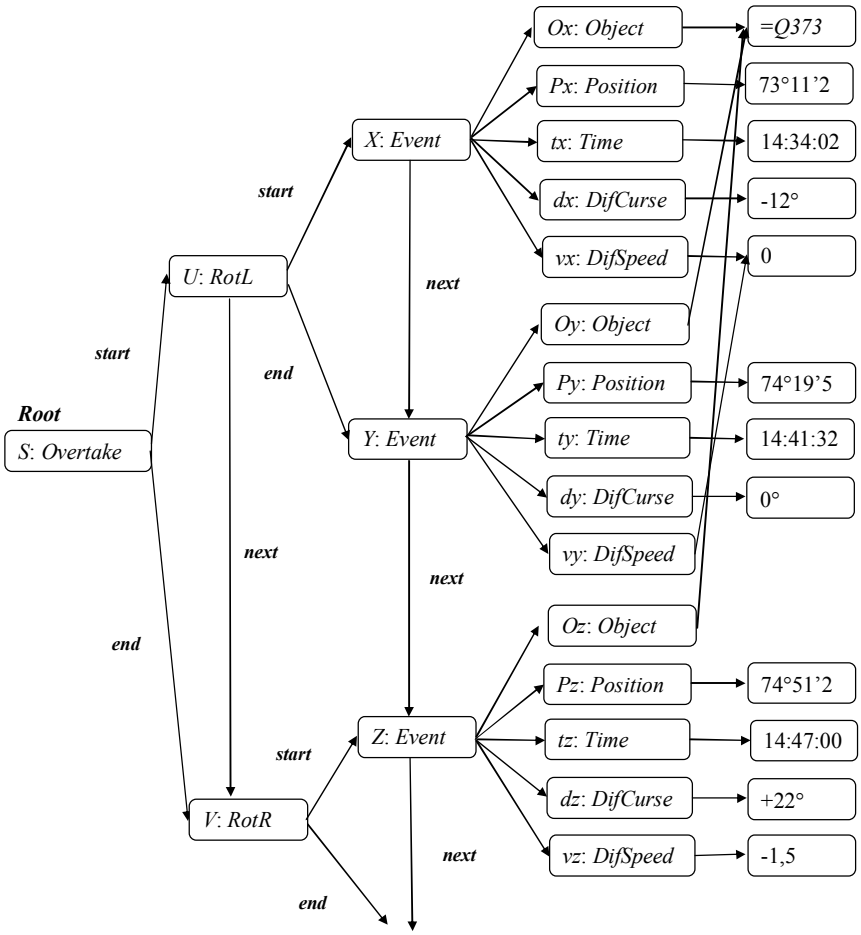
где $\hat{x}_i$ – имя параметра события, причем $x_1, ... x_n \in X_\psi$;

$\mathcal{X}_i^{t_j}$ – терм, содержащий значение параметра x_i в момент времени $t_j \in T$;

$t_j, t_{j+1}, ... t_{j+m}$ – последовательность моментов времени наблюдения значений параметра x_i.

Термом $\mathcal{X}_i$ будем называть выражение некоторого языка $\mathbf{\Lambda}$, являющееся формальным именем объекта-значения параметра x_i.

На рис. 1 представлена графическая иллюстрация записи терма ψ, где ψ указывает на основу (класс *overtake*) начальной (корневой) переменной $S \in \upsilon$, символьные имена классов заданы в $C \in E.\mathfrak{Z}$, символьные имена параметров заданы в $E.\mathfrak{Z}.X$, а имена переменных модели (метки) – на $E.\upsilon$.



$$\psi ::= S : overtake\left[start \to U : RotL\left[start \to X : Event\left[Ox \to Q373, Px \to 73^{\circ}11'27'', ...\right],\right.\right.$$
$$\left.\left.end \to Y : Event\left[Oy \to Q373, Py \to 74^{\circ}19'52'', ...\right]\right], end \to V : RotR[...]\right]$$

Рис. 1. Графическая иллюстрация представления информации о событиях

В событийной модели E формализация терма зависит от выбора языка Λ. Так, формализация терма на основе [14] дает возможность присваивать параметрам событий значения переменных, ссылки на переменные модели E, ссылки на связанные события того же или другого уровня абстракции, формулы языка Λ и т.д.

Терм также может содержать символ $\perp$, представляющий неизвестное либо недоступное значение параметра.

Таким образом, событие ψ является сложным объектом, принадлежащим некоторому классу событий c и включающим множество параметров количественной либо качественной природы

$x_1,\ldots x_n$, имеющих различную интерпретацию и степень точности, и представленных множеством слотов $\mathfrak{X}$.

Введем в рассмотрение функцию $\mathbf{root}(\psi)$, возвращающую *основу* (метку Y) события ψ.

Не имеющее параметров событие ψ, основа которого не определена, т.е. $n(\psi)=0 \wedge \mathbf{root}(\psi)=\perp$, назовем *пустым событием* и обозначим $\psi_\perp$.

Для формализации различного рода отношений, устанавливаемых между событиями и переменными в событийной модели E, введем понятие пути.

Путем $\rho(\nu,\mathfrak{x}_i)$ в модели E называется последовательность термов, ведущая по цепочке от переменной ν к терму $\mathfrak{x}_i$.

Два пути $\rho(\nu_1,\mathfrak{x}_i)$ и $\rho(\nu_2,\mathfrak{x}_i)$ называются *эквивалентными*, если они приводят к одному и тому же значению одного и того же терма $\mathfrak{x}_i$, т.е. $\rho(\nu_1,\mathfrak{x}_i)=\rho(\nu_2,\mathfrak{x}_i)$.

На рис. 2 показаны эквивалентные пути $\rho(Y,m5)$ и $\rho(X,m5)$, приводящие от различающихся переменных X и Y к одному значению $m5 \square Q373$.

Событие ψ *включает* другое событие ψ' в событийной модели E (обозначается как $\psi \sqsubseteq \psi'$), если:

1. основа события ψ' является подклассом основы события ψ, т.е. $\mathbf{root}(\psi) \preceq_1 \mathbf{root}(\psi')$;

2. всякий параметр, определенный для события ψ, определен

$$\psi ::= S:overtake\left[\begin{array}{l} start \sqsubset U:RotL\left[\begin{array}{l} start \sqsubset X:Event\left[\begin{array}{l} Ox \sqsubset m5:Q373 \\ Px \sqsubset m6:73^{\circ}11'27'' \\ tx \sqsubset m7:14:34:02 \\ dx \sqsubset m8:-12^{\circ} \\ vx \sqsubset m9:0 \end{array}\right] \\ end \sqsubset Y:Event\left[\begin{array}{l} Oy \sqsubset m5 \\ Py \sqsubset m10:74^{\circ}19'52'' \\ ty \sqsubset m11:14:41:32 \\ dy \sqsubset m12:0 \\ vy \sqsubset m9 \end{array}\right] \end{array}\right] \\ end \sqsubset V:RotR\left[\begin{array}{l} start \sqsubset Z:Event\left[\begin{array}{l} Oz \sqsubset m5 \\ Pz \sqsubset m13:74^{\circ}51'26'' \\ tz \sqsubset m14:14:47:00 \\ dz \sqsubset +22^{\circ} \\ vz \sqsubset -1,5 \end{array}\right] \\ end \sqsubset ... \end{array}\right] \end{array}\right]$$

Рис. 2. Пример записи о событии в виде терма

также (но может иметь иное значение) и для события ψ', т.е.

$$\forall x \in E.\mathcal{Z}.X:\ \psi.x \neq \perp\ \Rightarrow\ \psi'.x \neq \perp;$$

3. значение некоторого параметра, определенного для события ψ', является конкретизацией соответствующего параметра, определенного для события ψ, т. е.

$$\forall x \in E.\mathcal{Z}.X:\ \psi.x = v \neq \perp,\ \psi'.x = v' \neq \perp\ \Rightarrow\ v \sqsubseteq v';$$

4. обеспечивается эквивалентность путей следующего вида:

$$\rho\left(\mathbf{root}\left(\psi\right), x_1\right) = \rho\left(\mathbf{root}\left(\psi\right), x_2\right) = v\ \Rightarrow$$

$$\rho\left(\mathbf{root}\left(\psi'\right), x_1\right) = \rho\left(\mathbf{root}\left(\psi'\right), x_2\right) = v'\ \wedge\ \left(v \sqsubseteq v'\right)$$

Включение (*вложение*) событий является ключевой концепцией событийной модели E, позволяющей формализовать потоки событий и их отношения.

Потоки событий и классов событий

Потоком событий $\vec{S}$ в модели событий E будем называть упорядоченную относительно $<_T$ совокупность событий вида:

$$\vec{S} = \left[\psi_1, \psi_2, ...\psi_n\right], \qquad (6)$$

такую, что для всех событий совокупности $\psi_1.t \leq_T \psi_2.t \leq_T ... \leq_T \psi_n.t$.

j-е относительно $<_T$ событие ψ потока событий $\vec{S}_i \in E$ будем обозначать в дальнейшем как ψ_j^i, причем $\psi_i.t = t_j \in T$. В тех случаях, когда рассматривается единственный поток событий $\vec{S}$, будем использовать запись без надстрочного индекса вида ψ_j.

Полное упорядочение событийной модели E по $<_T$ при наличии $\langle t_0, T, <_T \rangle$ является *необходимым условием* ее адекватности.

Примем, что *длина потока событий* $\vec{S}$ (обозначается как $|\vec{S}|$) определяется его мощностью (n), т.е. $|\vec{S}| = n$.

Префиксом длины l потока событий $\vec{S} = [\psi_1, \psi_2, ...\psi_n]$ назовем последовательность событий вида $[\psi_1, \psi_2, ...\psi_l]$, где $l \le n$.

Суффиксом длины l потока событий $\vec{S} = [\psi_1, \psi_2, ...\psi_n]$ назовем последовательность событий вида $[\psi_{n-l}, ...\psi_{n-1}, \psi_n]$, где $l \le n$.

Любые два потока событий могут быть связаны между собой посредством оператора композиции $\circ$.

Потоком классов событий $\vec{V}$ в модели событий E будем называть упорядоченную совокупность классов событий вида:

$$\vec{V} = [c_1, c_2, ...c_n], \tag{7}$$

такую, что $\vec{V}_i \in E.\mathcal{Z}.I_1$.

Примем, что *длина потока классов* $\vec{V}$ (обозначается как $|\vec{V}|$) определяется его мощностью (n).

В построенной нами модели событий E для *любого* заданного потока событий $\vec{S}$ может быть определен соответствующий ему поток классов $\vec{V}$, такой что $\vec{V}_i = \mathbf{root}(\vec{S}_i)$.

Соотношение двух потоков событий

Пусть в модели событий E определены:

- потоки классов событий $\vec{U} = [u_1, u_2, ...u_n]$, $\vec{V} = [v_1, v_2, ...v_m]$, $\forall i \ \ u_i \in E.\mathcal{Z}.I_1$, $\forall j \ \ v_j \in E.\mathcal{Z}.I_1$, $n = |\vec{U}|$, $m = |\vec{V}|$, $m \le n$;

- потоки событий $\vec{S} = [\psi_1, \psi_2, ...\psi_k]$ и $\vec{R} = [\psi_1', \psi_2', ...\psi_l']$, $k = |\vec{S}|$, $l = |\vec{R}|$.

Поток классов событий $\vec{U}$ ***вложен*** в поток классов событий $\vec{V}$ (обозначим $\vec{U} \sqsubseteq \vec{V}$), если для всех $m \le n$ справедливо $u_1 \prec v_1, u_2 \prec v_2, ...u_m \prec v_m$.

Поток событий $\vec{S}$ ***вложен*** в поток событий $\vec{R}$ (обозначим $\vec{S} \sqsubseteq \vec{R}$), если для соответствующих им потоков классов событий $\vec{U} = \mathbf{root}\left(\vec{S}\right)$ и $\vec{V} = \mathbf{root}\left(\vec{R}\right)$ выполняется условие $\vec{U} \sqsubseteq \vec{V}$.

Поток событий $\vec{S}$ ***строго вложен*** в поток событий $\vec{R}$ (обозначим $\vec{S} \sqsubset \vec{R}$), если для всех $l \le k$ справедливо $\psi_1 \sqsubseteq \psi'_1, \psi_2 \sqsubseteq \psi'_2, ...\psi_l \sqsubseteq \psi'_l$.

Поток событий $\vec{R}$ ***включает*** поток $\vec{S}$, если $\vec{S} \sqsubseteq \vec{R}$.

Введем понятия сегментации и совмещения потоков событий, необходимые для сопоставления потоков событий, либо их фрагментов в случае неполного (частичного) совпадения.

Сегментацией $\mathbf{S}\left(\vec{S}, m\right)$ степени m потока событий $\vec{S} = \left[\psi_1, \psi_2, ...\psi_n\right]$ длины n будем называть последовательность из $m+1$ точек разрыва в диапазоне $\left[1, n\right]$, таких что:

$$\begin{gathered} \mathbf{S}\left(\vec{S}, m\right) = \left[s_1, s_2, ...s_m, s_{m+1}\right], \\ 1 = s_1 < s_2 < ... < s_m < s_{m+1} = n+1. \end{gathered} \tag{8}$$

Очевидно, что m-сегментация приводит к разбиению потока $\vec{S}$ на m сегментов $\left[\vec{S}^1, \vec{S}^2, ...\vec{S}^m\right]$ таким образом, что:

$$\begin{gathered} \left[\psi_1, ..., \psi_{s_2-1}\right] \circ \left[\psi_{s_2}, ..., \psi_{s_3-1}\right] \circ ... \\ ... \circ \left[\psi_{s_m}, ..., \psi_n\right] = \left[\left[\psi_{s_i}, ..., \psi_{s_{i+1}}\right]\right]_{i=1}^{m}, \end{gathered} \tag{9}$$

причем $\Sigma_{i=1}^{m} \left|S^i\right| = n$.

В общем случае, для любого заданного потока событий может быть определено конечное множество возможных сегментаций.

Совмещением потоков $\vec{S}$ и $\vec{R}$ назовем пару $\left\langle \vec{S}', \vec{R}' \right\rangle$, полученную вставкой пустых событий $\psi_\perp$ в оба потока таким образом, что $\left|\vec{S}'\right| = \left|\vec{R}'\right|$ и $\forall i$, $1 \le i \le \left|\vec{S}'\right|$, элемент $\vec{S}'\left[i\right]$ совмещен с элементом $\vec{R}'\left[i\right]$ при $\vec{S}'\left[i\right] \ne \perp$ и $\vec{R}'\left[i\right] \ne \perp$.

Формирование сегментации потоков событий

Отношения совмещения и сегментации потоков событий могут быть использованы для оценки подобия двух потоков событий.

Пусть $\vec{S}$ представляет собой наблюдаемый поток событий, а $\vec{R}$ – эталонный поток событий.

Рассматривая $\vec{S}$ и $\vec{R}$ как последовательности объектов (событий), установим, что различия между ними сводятся к:

- структурным различиям последовательности объектов;
- различиям индивидуальных свойств объектов, занимающих одинаковые позиции в обоих потоках.

Соответственно, можно рассматривать *подобие структуры* и *подобие состава* потоков событий.

Рассмотрим структурные различия потоков событий, для чего произведем k-сегментацию потока $\vec{S}$ и соответствующую l-сегментацию потока $\vec{R}$ согласно (9) (рис. 3).

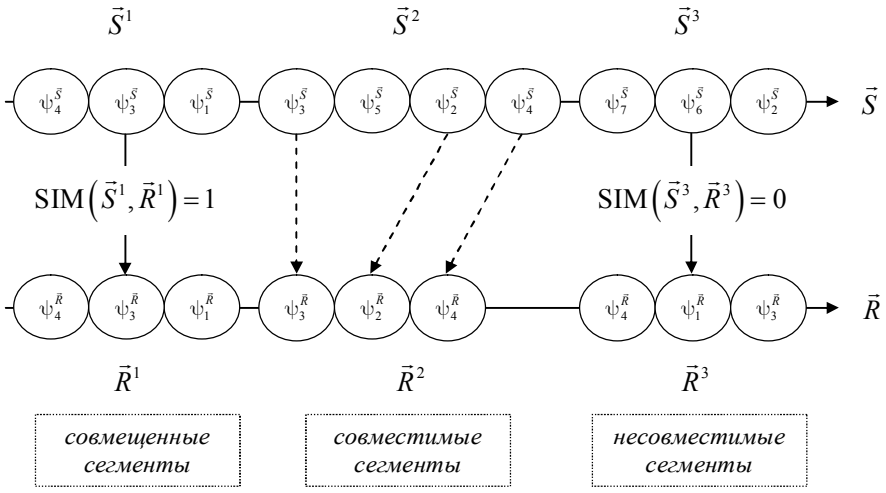


Рис. 3. Сегментация наблюдаемого и эталонного потоков событий

Обозначим символом $\sim$ сходство сегментов, а символом $\nsim$ – различие по их последовательной структуре.

Основываясь на [15], сопоставляемые потоки событий $\vec{S}$ и $\vec{R}$ можно представить как совокупности сегментов $\vec{S}^k$ и $\vec{R}^l$ трех типов:

а) *совмещенных* ($\vec{S}^1$ и $\vec{R}^1$ на рис. 3), если сопоставимые элементы (события) занимают одинаковые позиции в обоих сегментах, при

этом сопоставимые элементы могут иметь различный уровень абстракции:

$$\left\{\left(\vec{S}^k[i] \sqsubseteq \vec{R}^l[i]\right)\right\}_{0<i<\min\left(\left|\vec{S}^k\right|,\left|\vec{R}^l\right|\right)} : \ \vec{S}^k \sim_C \vec{R}^l; \tag{10}$$

б) *совместимых* ($\vec{S}^2$ и $\vec{R}^2$ на рис. 3), если сопоставимые элементы встречаются в обоих сегментах, но в различных позициях:

$$\left\{\left(\vec{S}^k[i] \sqsubseteq \vec{R}^l[j]\right)\right\}_{0<i\le\left|\vec{S}^k\right|,0<j\le\left|\vec{R}^l\right|,i\neq j} : \ \exists\left(\vec{S}^k\right) \to \widehat{S}^k \sim_A \vec{R}^l, \tag{11}$$

в этом случае сегменты могут быть совмещены с помощью оператора преобразования $\exists$ за конечное число шагов $q < \left|\vec{R}^l\right|$.

в) *несовместимых* ($\vec{S}^3$ и $\vec{R}^3$ на рис. 3), если элементы, присутствующие в сегменте одного потока, отсутствуют в сегменте другого

$$\begin{aligned} &\left\{\left(\vec{S}^k[i]\right)\right\}_{0<i\le\left|\vec{S}^k\right|} : \forall j \ \vec{S}^k[i] \not\sqsubseteq \vec{R}^l[j] \ \cup \\ &\left\{\left(\vec{R}^l[j]\right)\right\}_{0<j\le\left|\vec{R}^l\right|} : \forall i \ \vec{R}^l[j] \not\sqsubseteq \vec{S}^k[i] : \ \vec{S}^k \nsim \vec{R}^l \end{aligned}, \tag{12}$$

причем совместить сегменты за конечное число шагов q не представляется возможным.

Введем оценку структурного подобия сегментов $\vec{S}^k$ и $\vec{R}^l$ заданных потоков событий $\vec{S}$ и $\vec{R}$, и обозначим ее как $\mathbf{SIM}_{\ltimes}\left(\vec{S}^k, \vec{R}^l\right)$.

Безусловно, искать структурное подобие для несовместимых сегментов не имеет смысла, поэтому $\vec{S}^k \nsim \vec{R}^l : \ \mathbf{SIM}_{\ltimes}\left(\vec{S}^k, \vec{R}^l\right) = 0$.

Совмещенные сегменты однозначно эквивалентны по своей структуре, поэтому $\vec{S}^k \sim_C \vec{R}^l : \ \mathbf{SIM}_{\ltimes}\left(\vec{S}^k, \vec{R}^l\right) = 1$.

Совместимые сегменты для оценки структурного подобия должны быть совмещены посредством использования оператора $\exists$.

Совмещение совместимых сегментов потоков событий

Совмещение сегмента $\vec{S}^k$ наблюдаемого потока событий с сегментом $\vec{R}^l$ эталонного потока событий может производиться на основе известных способов редактирования [16] и выравнивания [17]. Способ выравнивания обычно применяется в предметных областях, связанных с биологией, геологией и т.д., где имеются цепочки объ-

ектов значительной длины, содержащие повторяющиеся блоки, например двойные спирали ДНК.

Поскольку для рассматриваемого класса задач события чаще выстраиваются в последовательности сегментов небольшой длины, предпочтительнее применить способ редактирования, суть которого состоит в поиске последовательности операций, преобразующих сегмент наблюдаемой последовательности таким образом, что структура сегмента преобразованной последовательности становится подобна структуре сегмента эталонной последовательности, что и приводит к совмещению сегментов [18].

Операцией преобразования op называется отображение i-го элемента исходной последовательности $\vec{S}^k$ в j-й элемент результирующей последовательности $\widehat{S}^k$, такое что $S^k[i] \xrightarrow{op} \widehat{S}^k[j]$.

Множество операций преобразования должно содержать минимум две основные операции:

- трехместную операцию *вставки события* ψ в сегмент S^k в позицию i: $Ins\left(S^k, \psi, i\right)$;

- двуместную операцию *отбрасывания события* из позиции i сегмента S^k: $Del\left(S^k, i\right)$.

С помощью вставок поток дополняется «пропущенными» событиями, с помощью отбрасывания фильтруются «шумовые» события. С помощью операции Ins могут быть вставлены не только конкретные или абстрактные, но и пустые события.

Оператором $\exists$ называется такая q-шаговая последовательность применения операций преобразования $\left[op_1, op_2, \ldots op_q\right]$ к сегменту исходной последовательности S^k наблюдаемого потока событий $\vec{S}$, что сегмент результирующей последовательности $\widehat{S}^k$ совмещается с сегментом R^l:

$$\exists\left(S^k\right) \xrightarrow{q} \widehat{S}^k \;:\; \widehat{S}^k \sim_C R^l .$$

На рис. 4 показано совмещение сегмента потока $\vec{S}$ с сегментом потока $\vec{R}$ методом редактирования, достигаемое использованием четырехшагового оператора преобразования

$$\exists(S) \xrightarrow{q=4} \widehat{S} = \left[Del(S,3), Del(S,7), Ins(S,\psi_7,9), Ins(S,\psi_{10},12) \right] ,$$

в результате которых получен такой сегмент $\widehat{S}$, что $\widehat{S} \sim_C R$.

Поскольку необходимо получить числовую оценку структурного подобия сегментов, процесс совмещения должен оцениваться.

Введем абсолютные значения *цены операции* вставки $F^{Ins\left(S^k,\psi,i\right)}$ и

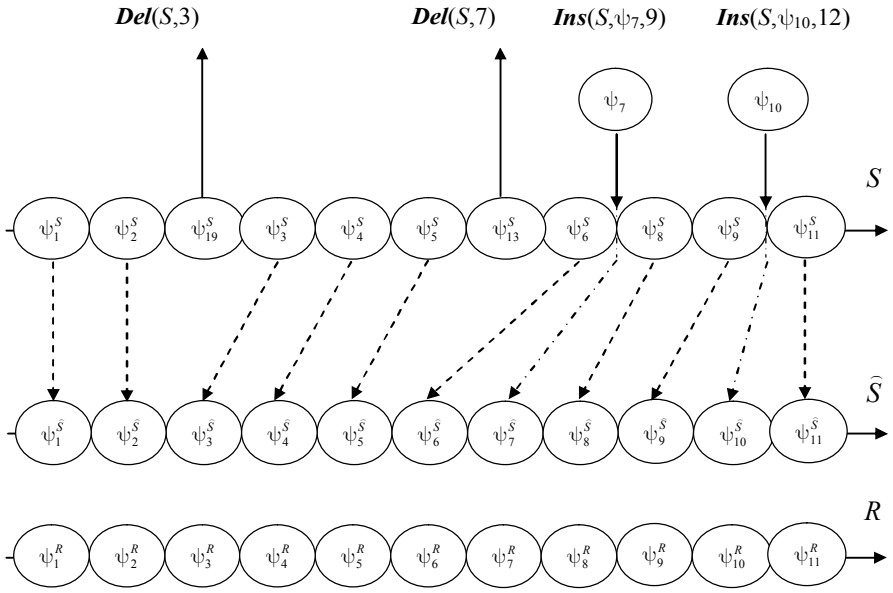


Рис. 4. Совмещение сегментов потоков событий $\vec{S}$ и $\vec{R}$ оператором

$$\exists\left(S^k\right) \xrightarrow{\left[Del(S,3),Del(S,7),Ins(S,\psi_7,9),Ins(S,\psi_{10},12)\right]} \widehat{S}^k$$

соответственно отбрасывания $F^{Del\left(S^k,i\right)}$ как

$$F^{Ins\left(S^k,\psi,i\right)} = \alpha \,,\; F^{Del\left(S^k,i\right)} = \beta \,. \qquad (13)$$

Следует отметить, что (13) можно усложнить, учитывая, например, вес позиции события-операнда в таксономической иерархии, как в [19], или статистические оценки относительного числа вхождений события-операнда в класс, которому принадлежит совмещаемый сегмент [20], однако, для нас критична минимальная вычислительная сложность.

Стоимость совмещения сегмента потока $\vec{S}$ с сегментом потока $\vec{R}$ – это сумма абсолютных значений цен операций преобразования

$\left[op_1, op_2, ...op_q\right]$, q-шаговая последовательность выполнения которых над $\vec{S}$ приводит в результате к получению такого сегмента потока $\widehat{S}$, что $\widehat{S} \sim_C R$:

$$F^{\exists=\left[op_1, op_2, ...op_q\right]} = \sum_{i=1}^{q} F\left(op_q\right). \qquad (14)$$

Для обеспечения линейной оценки по вычислительной сложности удобно принять $\alpha = \beta = 1$.

Оценка подобия сегментов потоков событий

Подобие сегмента $\vec{S}^k$ наблюдаемого потока событий сегменту $\vec{R}^l$ эталонного потока событий рассмотрим с двух позиций – подобия структуры сегмента $\vec{S}^k \sim_C \vec{R}^l$ и подобия состава сегмента $\vec{S}^k \sim_E \vec{R}^l$.

Оценка подобия структуры сегментов $\vec{S}^k$ и $\vec{R}^l$ (обозначим как $\mathbf{SIM}_{\ltimes}\left(\vec{S}^k, \vec{R}^l\right)$) определяется следующим образом:

$$\mathbf{SIM}_{\ltimes}\left(\vec{S}^k, \vec{R}^l\right) = \begin{cases} 0, \text{ если } \vec{S}^k \nsim \vec{R}^l \\ 1, \text{ если } \vec{S}^k \sim_C \vec{R}^l \\ 1 - \dfrac{F^{\exists}}{\left|\vec{R}^l\right|}, \text{ если } \exists\left(S\right) \overset{q}{\to} \widehat{S} \ \wedge \widehat{S} \sim_C R \end{cases}. \qquad (15)$$

Очевидно, что в случае использования цен операций $\alpha = \beta = 1$ уже при $q = \left|\vec{R}^l\right|$ совмещение сегментов $\vec{S}^k$ и $\vec{R}^l$ не имеет смысла, поскольку в этом случае $\mathbf{SIM}_{\ltimes}\left(\vec{S}^k, \vec{R}^l\right) = 0$.

Перейдем к оценке подобия событий, составляющих сегмент потока. Для того, чтобы избавиться от необходимости перебора и сравнения значений всех параметров, что приводит к экспоненциальной по мощности множества параметров оценке сложности [21], можно воспользоваться имеющейся в событийной модели E таксономической иерархией $\left\langle \perp_1, C, \prec_1 \right\rangle$.

Очевидно, что чем более абстрактную позицию в C занимает одно событие относительно другого, тем меньше должна быть оценка подобия событий.

Пусть $\hat{C}$ – вершина таксономической иерархии $\langle \perp_1, C, \prec_1 \rangle$, $\|C\|$ – ее глубина. Для того, чтобы оценить разность между уровнем двух событий в таксономической иерархии $\langle \perp_1, C, \prec_1 \rangle$, необходимо вычислить длину пути ρ, ведущего из вершины иерархии к классу $c = \mathbf{root}(\psi)$ каждого события и соотнести соответствующие значения.

Оценка подобия событий, составляющих сегменты $\vec{S}^k$ и $\vec{R}^l$ (обозначим как $SIM\left(\vec{S}^k[i], \vec{R}^l[i]\right)$) определяется отношением уровней положения классов $\mathbf{root}\left(\vec{S}^k[i]\right)$, $\mathbf{root}\left(\vec{R}^l[i]\right)$ в иерархии $\langle \perp_1, C, \prec_1 \rangle$:

$$SIM\left(\vec{S}^k[i], \vec{R}^l[i]\right) = 1 - \frac{\left|\rho\left(\hat{C}, \mathbf{root}\left(\vec{S}^k[i]\right)\right)\right| - \left|\rho\left(\hat{C}, \mathbf{root}\left(\vec{R}^l[i]\right)\right)\right|}{\|C\|}. \qquad (16)$$

Оценка подобия состава сегментов $\vec{S}^k$ и $\vec{R}^l$ (обозначим как $\mathbf{SIM}_\star\left(\vec{S}^k, \vec{R}^l\right)$) определяется как среднее арифметическое оценок подобия событий, составляющих соответствующие сегменты:

$$\mathbf{SIM}_\star\left(\vec{S}^k, \vec{R}^l\right) = \frac{\sum_{i=1}^{\left|\vec{S}^k\right|} SIM\left(\vec{S}^k[i], \vec{R}^l[i]\right)}{\left|\vec{S}^k\right|}. \qquad (17)$$

Тогда оценка подобия сегментов $\vec{S}^k$ и $\vec{R}^l$ (обозначаемая как $\mathbf{SIM}\left(\vec{S}^k, \vec{R}^l\right)$) определяется произведением оценки подобия структуры и оценки подобия состава сегментов:

$$\mathbf{SIM}\left(\vec{S}^k, \vec{R}^l\right) = \mathbf{SIM}_\star\left(\vec{S}^k, \vec{R}^l\right) \cdot \mathbf{SIM}_\times\left(\vec{S}^k, \vec{R}^l\right). \qquad (18)$$

Если известны оценки подобия всех сегментов, составляющих потоки событий, можно определить оценку подобия собственно потоков событий.

Динамическая оценка подобия двух потоков событий

При оценке подобия совместимых сегментов производится поиск возможных совмещений суффикса наблюдаемого потока событий $\vec{S}$ с префиксом эталонного потока событий $\vec{R}$.

Динамическая оценка подобия потоков $\vec{S}$ и $\vec{R}$, обозначаемая как $\mathbf{SIM}\left(\vec{S},\vec{R}\right)$, может быть получена по принципу максимально возможного совмещения. При этом, если в $\vec{S}$ существует сегмент $\vec{S}^k$, несовместимый с соответствующим сегментом потока $\vec{R}$, в последнем для него создается дополняющий сегмент $\vec{R}^l$.

Дополняющим сегментом для сегмента $\vec{S}^k$ назовем такой сегмент $\vec{R}^l$, что $\left|\vec{S}^k\right|=\left|\vec{R}^l\right|$ и $\underset{l\le\left|\vec{R}^l\right|}{\forall l}\quad \vec{R}^l[l]=\psi_\perp$.

Существуют два способа оценки подобия потоков событий на основе оценок подобия сегментов:

- аддитивная оценка $\mathbf{SIM}\left(\vec{S},\vec{R}\right)=\sum_{i.j=1}^{l}\mathbf{SIM}\left(\vec{S}^i,\vec{R}^j\right)$;

- мультипликативная оценка $\mathbf{SIM}\left(\vec{S},\vec{R}\right)=\prod_{i,j=1}^{l}\mathbf{SIM}\left(\vec{S}^i,\vec{R}^j\right)$.

Если принять за основу мультипликативную оценку, присутствие во входном потоке $\vec{S}$ несовместимых относительно $\vec{R}$ сегментов приведет в конечном итоге к $\mathbf{SIM}\left(\vec{S},\vec{R}\right)=0$, что может вывести из рассмотрения некоторые не вполне подобные текущей ситуации, но вполне уместные для принятия решений прецеденты.

Вычисление $\mathbf{SIM}\left(\vec{S},\vec{R}\right)$ необходимо производить таким образом, чтобы наличие в потоках несовместимых сегментов не приводило к обнулению оценки подобия потоков.

Динамическая оценка подобия двух потоков событий $\vec{S}$ и $\vec{R}$ определяется как среднее арифметическое максимальных оценок подобия составляющих сегментов:

$$\mathbf{SIM}_{\mathcal{D}}\left(\vec{S},\vec{R}\right)=\frac{\sum_{j=1}^{l}\max_{1<i<k}\mathbf{SIM}\left(\vec{S}^i,\vec{R}^j\right)}{l}. \qquad (19)$$

Как видно из (19), $\mathbf{SIM}_{\mathcal{D}}\left(\vec{S},\vec{R}\right)$ является динамической оценкой, поскольку появление каждого нового события $\psi_i^{\vec{S}}$ в наблюдаемом потоке событий $\vec{S}$ приводит к изменению суффиксного сегмента $\vec{S}^k$, в результате чего, как минимум, пересчитывается оценка подобия сегмента $\vec{S}^k$, которая в свою очередь влияет на оценку подобия потоков $\vec{S}$ и $\vec{R}$.

В других ситуациях получение нового события может изменить k-сегментацию потока $\vec{S}$, от чего изменятся оценки подобия для ряда сегментов, что также изменит оценку подобия потоков $\vec{S}$ и $\vec{R}$.

Полученную оценку подобия потоков событий необходимо нормировать, т. е. привести к числовому диапазону $[0,1]$, ограничив максимально возможной величиной подобия (например, подобие эталонного потока событий самому себе, поскольку $\mathbf{SIM}\left(\vec{R},\vec{R}\right)=1$).

Нормализованная динамическая оценка подобия потоков событий $\mathbf{SIM}\left(\vec{S},\vec{R}\right)$ представляет собой отношение оценки подобия потоков событий $\vec{S}$ и $\vec{R}$ друг другу к оценке подобия эталонного потока событий $\vec{R}$ самому себе:

$$\mathbf{SIM}\left(\vec{S},\vec{R}\right)=\mathbf{SIM}_{\mathcal{D}}\left(\vec{S},\vec{R}\right)\Big/\mathbf{SIM}_{\mathcal{D}}\left(\vec{R},\vec{R}\right). \qquad (20)$$

Предложенный метод оценки подобия потоков событий, основанный на принципе максимально возможного совмещения, имеет низкую оценку вычислительной сложности $O\left(m\times n\right)$, где $m=\left|\vec{S}\right|, n=\left|\vec{R}\right|$, и зависит исключительно от длин сравниваемых потоков, что допускает его использование в прецедентных системах реального времени.

Выводы

В работе получили дальнейшее развитие теоретические основы динамических прецедентных систем реального времени.

Построена расширенная модель событий, в основу которой положены множество иерархий и явная временная шкала. Модель допускает события, характеризуемые неполными и недостоверными векторами параметров. Формализовано понятие потока событий,

введены отношения вложения и включения потоков событий. Определены отношения сегментации и совмещения потоков, применимые для сопоставления двух потоков событий либо их фрагментов в случае неполного (частичного) совпадения.

Предложен метод динамической оценки подобия потоков событий, основанный на принципах максимально возможного совмещения сегментов потоков способом редактирования, таксономической оценки подобия событий и стоимостной оценки подобия сегментов. Метод работоспособен в условиях неполной и неточной информации, а низкая оценка его вычислительной сложности делает его применимым для реализации в системах реального времени.

Литература

1. *Павлов В.В.*, *Антомонов Ю.Г.*, *Ивахненко А.Г.* Эргатические динамические системы управления. К.: Наукова думка, 1975. 160 с.
2. *Hellerstein J.L.* Discovering Actionable Patterns in Event Data // IBM Systems Journal. 2002. Vol. 41. №3. Pp.475-492.
3. *Aamodt A.* Case-based reasoning: foundational issues, methodological variations, and system approaches // AI Comm. 1994. Vol. 7. №1. Pp.39-59.
4. *Шерстюк В.Г.* Основы теории динамических сценарно-прецедентных интеллектуальных систем. Херсон: ХНТУ, 2012. 432 с.
5. *Pal S.K.*, *Shiu S.C.* Foundations of Soft Case-Based Reasoning. N.Y.: J.Wiley&Sons, 2004. 274 p.
6. *Tversky A.* Features of Similarity // Psychological Review. 1977. Vol.84. Pp.327-352.
7. *Leake D.* Case-based reasoning – experiences, lessons, and future directions. – N.Y.: AAAI Press-MIT Press, 1996. 318 p.
8. *Chen Z.* Analogical problem solving: a hierarchical analysis of procedural similarity // J. Exp. Psyc. Learning, Memory, Cognition. 2002. Vol.28. №1. Pp.81-98.
9. *Levenshtein V.I.* Binary codes capable of correcting deletions, insertions, and reversals // Cybernetics and Control Theory. 1966. Vol.10. №8. Pp.707-710.
10. *Schrodt P.A.* Pattern Recognition of International Crises using Hidden Markov Models // Political Complexity: Nonlinear Models of Politics. – University of Michigan Press, 2000. Pp.296-328.
11. *Gusfield D.* Algorithms on Strings, Trees, and Sequences. – Cambridge: Cambridge University Press Syndicate, 1997. 381 p.
12. *Keane M.T., Smyth B.* Dynamic Similarity: A Processing Perspective on Similarity // Similarity and Categorization. – Oxford: Oxford University Press, 2001. 296 p.
13. *Martin F.J.* Case-Based Sequence Analysis in Dynamic, Imprecise, and Adversarial Domains: tesi doctoral. – Barcelona: Universitat Politecnica De Catalunya, 2004. 285 p.
14. *Ait-Kaci H.* Description logic vs. order-sorted feature logic // Proc. of 2007 Int. Workshop on Description Logics (DL2007). – Bozen-Bolzano, Italy, 2007. Vol.250. Pp.147-154.

15. *Li Q., Feng L., Pei J.* Effective Similarity Analysis over Event Streams Based on Sharing Extent // Proc. of the J. Int. Conf. on Advances in Data and Web Management APWeb/WAIM'09. – Berlin: Springer-Verlag, 2009. Pp.308-319.

16. *Loshin D.* The Practitioner's Guide to Data Quality Improvement. – Burlington: Elsevier, Morgan Kaufmann, 2011. 432 p.

17. *Ortet P., Bastien O.* Where Does the Alignment Score Distribution Shape Come from? // Evolutionary Bioinformatics. 2010. Vol.6. Pp.159-187.

18. *Левенштейн В.И.* Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады АН СССР. 1965. Т.163. Вып.4. С.845–848.

19. *Datar M., Muthukrishnan S.* Estimating Rarity and Similarity over Data StreamWindows // Proc. of the 10th Annual European Symposium on Algorithms. – Springer-Verlag, 2002. Vol.2461. Pp.323-334.

20. *Hyvönen S., Gionis A., Mannila H.* Recurrent Predictive Models for Sequence Segmentation // Proc. of Int. Data Analysis Conf. (IDA-2007). – Springer-Verlag, 2007. Vol.4735. Pp.195-206.

21. *Li W., Godzik A.* Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences // Bioinformatics. 2006. Vol.22. Pp.1658-1659.

E-MAIL: V_SHERSTYUK@BIGMIR.NET