МЕТОД ВИЗУАЛИЗАЦИИ ЗОН НЕСТАБИЛЬНОСТИ В РЯДАХ ИЗМЕРЕНИЙ

Ландэ Д. В.,
Институт проблем регистрации информации НАН Украины

Визуализации особенностей рядов измерений посвящены многочисленные исследования. В частности, Фурье-, вейвлет- и фрактальный анализ [1-3] позволяют выявлять гармонические (периодические) составляющие, тренды, локальные особенности. Предлагаемый метод, основанный на алгоритме сглаживания пиковых значений и концепции клеточных автоматов (Smoothing, Cellular Automata – SCA), позволяет выявлять стационарные области, области с частыми (возможно небольшими по амплитуде) скачками значений, гармонические составляющие временного ряда. С помощью этого метода не детектируются абсолютные амплитудные всплески, однако SCA хорошо показал себя при детектировании зон нестабильности на «изрезанных» структурах данных, близких к фрактальным. К таким данным относятся, в частности, временные ряды, связанные с объемами публикаций в веб-пространстве по определенным тематикам, которые рассматриваются ниже как иллюстрация метода.

В предлагаемой модели каждому значению исходного ряда измерений $x_0(t)$ (обозначим исходный ряд как $X_0 = \{x_0(t)\}$) соответствует одна клетка клеточного автомата [4]. По ряду измерений строится сглаженный по приведенному ниже правилу ряд $X_1 = \{x_1(t)\}$. Затем ряду X_1 ставится в соответствие ряд X_2 (получаемый из X_1 по тому же алгоритму сглаживания) и т.д. Правило сглаживания пиков заключается в том, что значения, которые принимают

элементы рядов измерений $x_k(t) \in X_k$ (k – шаг сглаживания, t – номер элемента ряда измерений) составляют:

$$x_k(t) = \begin{cases} x_{k-1}(t), \ \ if \ \ x_{k-1}(t) \le \dfrac{x_{k-1}(t-1) + x_{k-1}(t+1)}{2}; \\ \dfrac{x_{k-1}(t-1) + x_{k-1}(t+1)}{2}, \ \ if \ \ x_{k-1}(t) > \dfrac{x_{k-1}(t-1) + x_{k-1}(t+1)}{2}. \end{cases}$$

Цвет клетки с номером t одномерной клеточной структуры, соответствующей X_k, белый, если $x_k(t)$ совпадает с $x_{k-1}(t)$, в противном случае – черный. Таким образом, каждой клетке соответствует значение $x_k(t)$ и значение ее цвета. (*Необходимо отметить, что такую систему нельзя считать каноническим клеточным автоматом, так как в общем случае клеткам может соответствовать бесконечное множество значений* $x_k(t)$ *и два значения цвета*).

Таким образом, алгоритм сглаживания пиков и визуализации можно представить в следующем виде:

Шаг 1. Задается значение предельного количества итераций N и исходному ряду измерений присваивается индекс $k = 0$.

Шаг 2. Индекс k увеличивается на 1. Рассчитываются значения сглаженного временного ряда $x_k(t)$ в соответствии приведенной выше формулой.

Шаг 3. Рассчитываются значения цветов соответствующих клеток для всех значений t, которые отображаются.

Шаг 4. Если текущий шаг итерации k не превышает заданное заранее количество итераций N или $\exists t: \ x_k(t) \ne x_{k-1}(t)$, то происходит переход к шагу 2. Иначе процедура завершается.

Рассмотрим результаты выполнения данного алгоритма для простейших структур, которые, как показывает практика, охватывает все возможные варианты визуализации.

Очевидно, если значения ряда измерений в рассматриваемой зоне представляют собой вогнутое (выпуколое вниз множество), то сразу же, на первой итерации получим $\forall t:\ x_1(t) = x_0(t)$ и выполнение алгоритма прерывается.

Если область значений представляют собой выпуклое вверх множество, то визуальное представление клеточных автоматов принимает вид сплошной черной полосы (рис. 1: вертикальная ось – номер шага итерации, а горизонтальная ось – номер элемента ряда измерений).

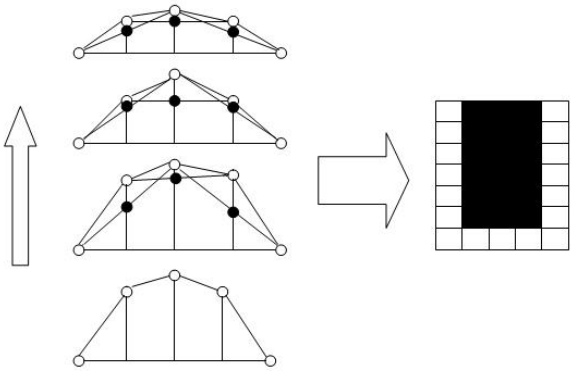
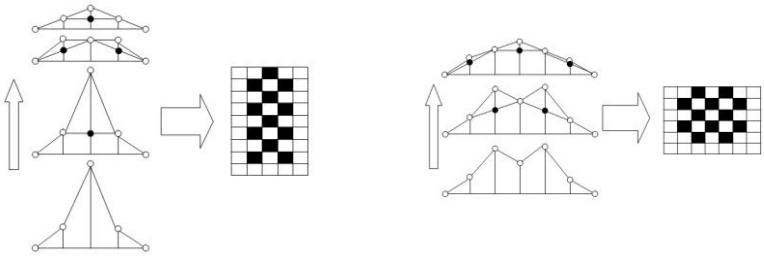


Рис. 1 – Выпуклое вверх множество точек

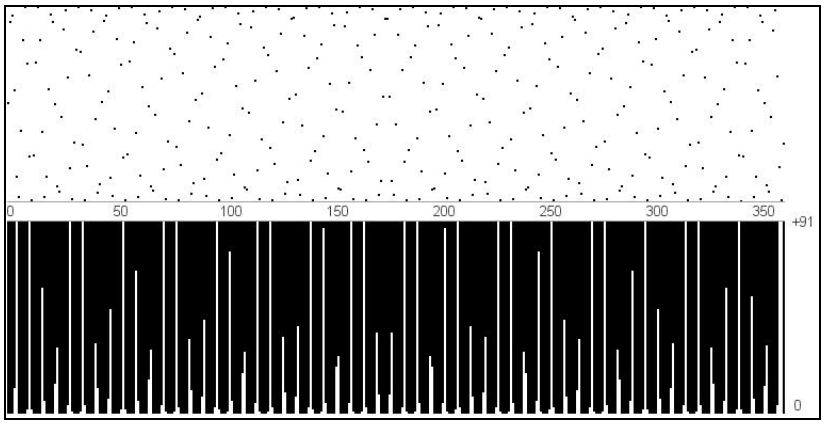
Единичные всплески значений в исходном ряде измерений (рис. 2а) и области изрезанности (рис. 2б) могут вызывать появление структур типа «шахматной доски».



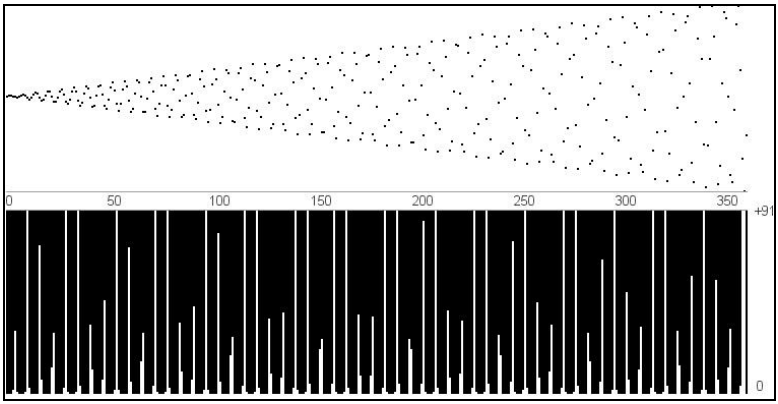
а б

Рис. 2 – Появление структур типа «шахматной доски»

Диаграммы, формируемые в результате визуализации в соответствии с предложенным алгоритмом, позволяют выявлять периодические составляющие, это можно продемонстрировать на примере двух функций $y = \sin(x)$ и $y = x\sin(x)$ (рис 3а и 3б, соответственно). На этих рисунках верхняя часть каждого из них – исходные данные, а нижняя – слои одномерного клеточного автомата (от $k = 0$ до $k = 91$).



а



б

Рис. 3 – Отображение простых периодических составляющих

Функции, содержащие несколько гармонических составляющих, позволяют визуально выделять из них. На рис. 4 приведен пример для ряда значений, соответствующего функции $y = \sin(x) + \cos(3x)$. Верхняя огибающая диаграммы SCA в данном случае соответствует функции $\cos(3x)$, а нижняя – функции $\sin(x)$.

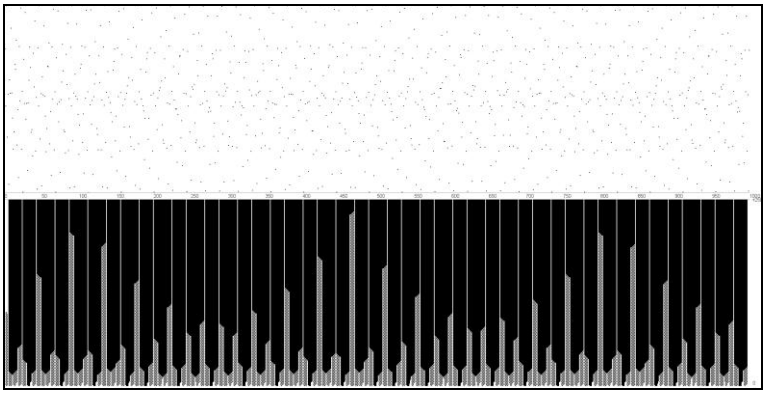


Рис. 4 – Отображение комплексных гармонических составляющих

На рис. 5. представлен интерфейс системы контент-мониторинга InfoStream [5], с помощью которой были получены данные – реальный временной ряд измерений, соответствующий посуточным объемам публикаций в веб-пространстве по некоторой заданной тематике (точки ряда – объемы публикаций за сутки) представлено на рис. 5.

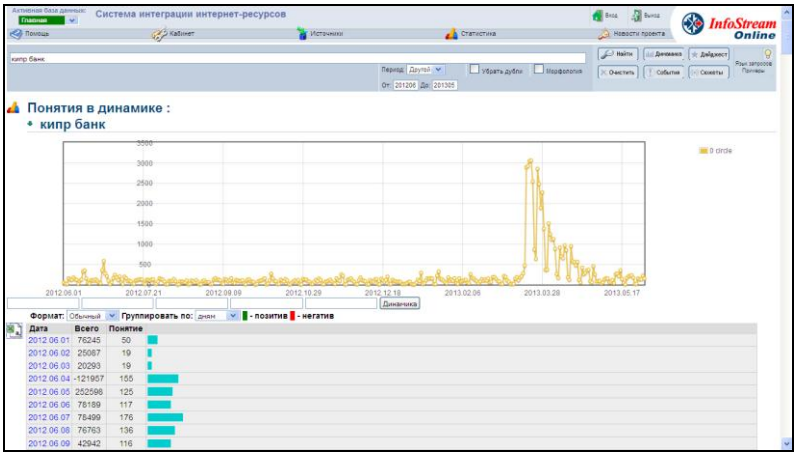


Рис. 5 – Интерфейс режима «Динамика публикаций» системы контент-мониторинга InfoStream

В качестве примера, на рис. 6 приведена визуализация динамики публикаций в RUNet (тематических информационных потоков) по запросу «Банк & Кипр» за период с июня 2012 г. по май 2013 г. Как видно, пик публикаций, связанных с банковским кризисом на Кипре приходится на 17-18 марта 2013 года. Кроме того четко отслеживаются недельные периодичности публикаций (минимумы – праздники, субботы и воскресенья) и периоды нестабильности.

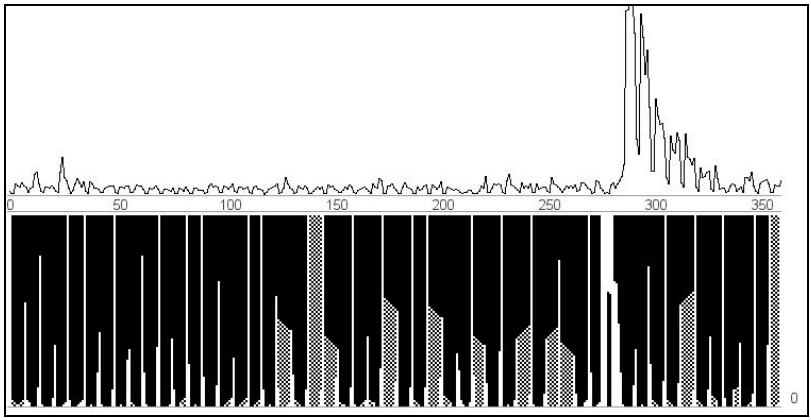


Рис. 6 – Отображение динамики тематического информационного потока по запросу «Банк & Кипр»

Как известно, тематика кипрского кризиса является частью более общей тематики, связанной с офшорами в банковской сфере, которые, безусловно, также находят свое отражение в веб-пространстве. На рис. 7 отражена динамика публикаций по запросу «Банк & Офшор» и соответствующая SCA-диаграмма. И если обычная взаимная корреляция временных рядов, соответствующих двум рассматриваемым тематическим информационным потокам составляет всего лишь 0,7, то визуальное сходство SCA-диаграмм очевидно.

Предложенный метод CSA является относительно простым в программной реализации и линейным по сложности, так как базируется на алгоритме сглаживания пиков и концепции клеточных автоматов. Он позволяет визуально выявлять единичные и нерегулярные «всплески», резкие колебания, скачки значений, зоны нестабильности количественных показателей в разные периоды времени. Метод CSA испытывался при анализе временных рядов, связанных с объемами публикаций в веб-пространстве по определенным темам [6].

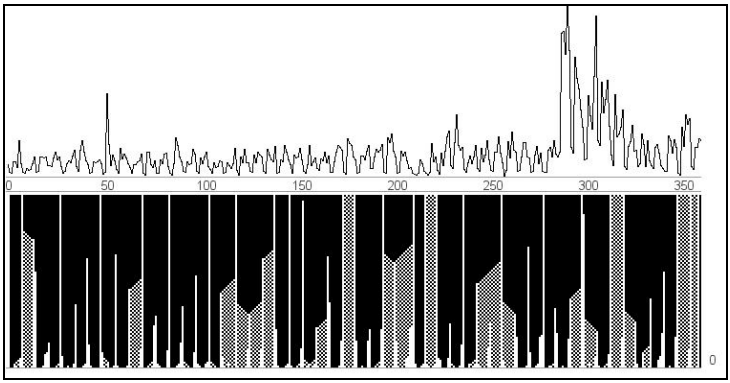


Рис. 7 – Отображение динамика тематического информационного потока по запросу «Банк & Офшор»

Также результаты анализа с помощью метода CSA сравнимы с результатами Фурье-анализа, применяемого к зашумленным периодическим последовательностям, и позволяют эффективно детектировать гармонические составляющие. Метод CSA, ввиду простоты реализации и наглядности, может эффективно применяться при анализе временных рядов в таких областях как экономика и социология.

Итак, приведенный метод может быть использован для визуализации неоднородностей в динамике информационных потоков, однако, проблема прогнозирования динамики остается открытой. Воспроизведение результатов во времени является серьезной проблемой при моделировании и оценке информационных процессов, составляет основу научной методологии. В настоящее время только ретроспективный анализ уже реализованных информационных потоков остается относительно надежным способом их верификации.

Литература

1. *Astafyeva N.M.* Wavelet-analysis: theoretical basis and application examples // Uspekhi Fizicheskih Nauk. – 1996. – **166**. – № 11. – P. 1145-1170.

2. *Buckheit J., Donoho D.* Wavelab and reproducible research // Stanford University Technical Report 474: Wavelets and Statistics Lecture Notes, 1995. – 27 p.

3. *Lande D.V., Snarskii A.A.* Diagram of measurement series elements deviation from local linear approximations // arXiv:0903.3328.

4. *Von Neumann J.* Theory of self-reproducing automata / Ed. A.W. Burks. – Urbana, University of Illinois Press, 1966. – 324 p.

5. *Григорьев А.Н., Ландэ Д.В. и др.* Мониторинг новостей из Интернет: технология, система, сервис: научно-методическое пособие. – К.: ООО «Старт-98», 2007. – 40 с.

6. *Lande D.V., Braichevskii S.M.* Dynamics of thematic information flows // arXiv:0805.4081.