УДК 681.3: 519.68

Метод визуального анализа временных рядов

Ландэ Д.В., д.т.н., с.н.с., Снарский А.А., д. ф.-м. н., проф., Зубок В.Ю.,
Национальный технический университет Украины «КПИ», г. Киев
dwl@visti.net

В работе предложен метод выявления трендов, периодичностей, локальных особенностей в рядах измерений (ΔL-метод), базирующийся на технологии DFA. Идея метода заключается в отображении значений абсолютного отклонения точек ряда накопления значений измерений от значений соответствующих им линейных аппроксимаций.

Введение

Задачи выявления и визуализации трендов, выявления гармонических составляющих, локальных особенностей временных рядов, фильтрации шума сегодня решаются методами фрактального, вейвлет- и Фурье-анализа. В данной работе предлагается метод визуализации, который базируется на технологии DFA (Detrended fluctuation analysis) и отличается своей универсальностью и вычислительной простотой.

Метод DFA

Метод DFA (Detrended fluctuation analysus) [1,2] чаще всего используется для выявления статистического самоподобия сигналов. Суть этого метода заключается в следующем. Пусть имеется ряд измерений x_t, $t \in 1,...,N$. Обозначим среднее значение этого ряда измерений: $\langle x \rangle$. Из исходного ряда строится ряд накопления: $X_t = \sum_{k=1}^{t} \left(x_k - \langle x \rangle \right)$, который затем разделяется на временные окна длиной L. Далее строится линейная аппроксимация по значениям $X_{k,j,L}$ из $X_{j,L}$ внутри каждого окна (в свою очередь, $X_{j,L}$ - подмножество X_t, $j = 1,...,J$, $J = N / L$ - количество окон наблюдения) и рассчитывается отклонение точек ряда накопления от линейной аппроксимации:

$$E(j,L)=\sqrt{\frac{1}{L}\sum_{k=1}^{L}\left(X_{k,j,L}-L_{k,j,L}\right)^2}=\sqrt{\frac{1}{L}\sum_{k=1}^{L}|\Delta_{k,j,L}|^2},$$

где $L_{k,j,L}$ - значение линейной аппроксимации в точке $t=(j-1)L+k$, $|\Delta_{k,j,L}|$ - абсолютное отклонение элемента $X_{k,j,L}$ от локальной линейной аппроксимации. Затем вычисляется среднее значение $F(L)=\frac{1}{J}\sum_{j=1}^{J}E(j,L)$, после чего, в случае $F(L)\propto L^{\alpha}$ (α -константа), делаются выводы о наличии статистического самоподобия, характеризуемого параметром α.

ΔL -метод

В отличие от рассмотренного выше метода DFA, рассмотрим поведение не усредненного значения $F(L)$, а абсолютного отклонения точек ряда накопления от линейной аппроксимации $|\Delta_{k,j,L}|$. Построение соответствующих диаграмм значений, зависящих от двух параметров - L и t, назовем ΔL-методом визуализации. Следует отметить, что разделение исходного интервала значений $t\in 1,...,N$ на J непересекающихся окон наблюдения приводит к некоторому «неравноправию» точек внутри этих окон, что не является принципиальным в случае последующей интегрированной оценки, но существенно при анализе локальных значений и визуализации. Поэтому в рамках ΔL-метода для каждой точки t выбирается такое окно наблюдения длиной L, чтобы данная точка оказывалась в его центре (или со смещением в 1 в случае четных L). Безусловно, с учетом этой поправки, замедляется скорость вычисления $|\Delta_{k,j,L}|$, что, однако, в значительной мере компенсируется простотой алгоритма.

Применение ΔL -метода

В качестве исследуемого временного ряда, на котором будем рассматриваться возможности метода, выберем ряд из посуточного количества публикаций сообщений по определенной тематике в веб-среде в течение 2008 года (рис. 1). Этот ряд получен с помощью системы контент-мониторинга InfoStream, регулярно сканирующей свыше 3000 веб-сайтов [3]. Чаще всего ряды, соответствующие тематическим

информационным потокам, обладают свойствами статистического самоподобия [4], что подтверждается, в частности, методом DFA (рис.2).

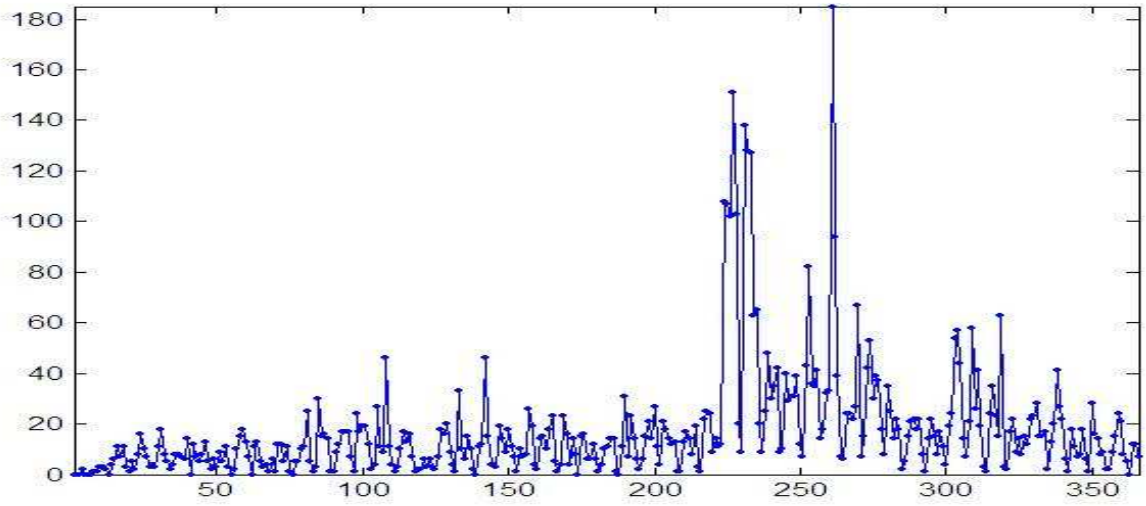

Рис. 1. Временной ряд интенсивностей публикаций по заданной тематике (ось абсцисс – дни года, ось ординат – количество публикаций)

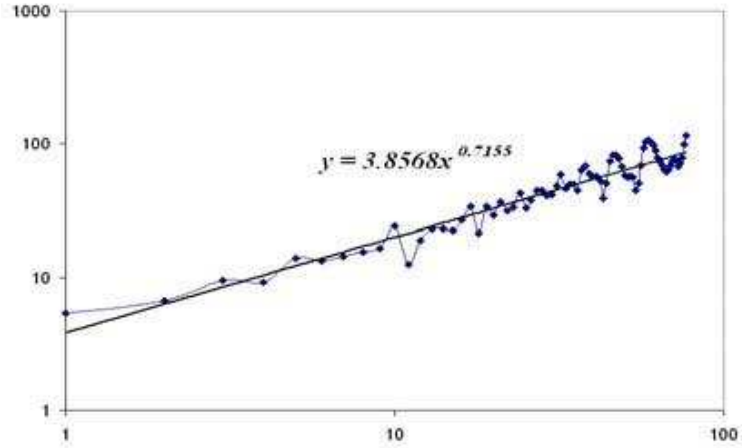

Рис. 2. Диаграмма значений $F(L)$ для рассматриваемого ряда измерений (ось абсцисс – величина окна наблюдения, ось ординат – значения показателя DFA)

«Рельефные диаграммы», получаемые в результате предложенного ΔL-метода (рис. 3), где более светлые тона соответствуют большим значениям $|\Delta_{k,j,L}|$, напоминают скейлограммы, получаемые в результате непрерывных вейвлет-преобразований.

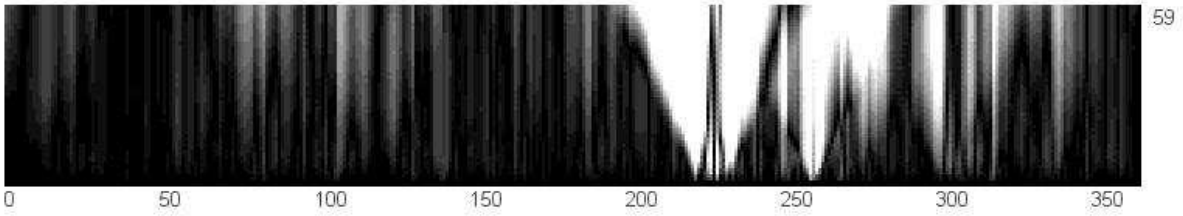

Рис. 3. ΔL-диаграмма временного ряда интенсивности тематических публикаций

(ось абсцисс – дни года, ось ординат – величина окна измерений)

Сравнение с вейвлет-анализом

ΔL -диаграммы внешне похожи на скейлограммы, получаемые в результате вейвлет-анализа. Основная идея вейвлет-преобразований состоит в том, что некоторый числовой ряд как и в рассмотренном выше методе разделяется на «окна наблюдения», на каждом из которых генерируется набор коэффициентов. По своей природе вейвлет-коэффициенты представляют собой определенную степень близости исследуемого ряда измерений с некоторой специальной функцией, называемой вейвлетом [5, 6].

Непрерывное вейвлет-преобразование для функции $f(t)$ строится с помощью непрерывных масштабных преобразований и переносов вейвлета $\psi(t)$ с произвольными значениями масштабного коэффициента a и параметра сдвига b:

$$W(a,b) = (f(t),\psi(t)) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t)\psi^* \left(\frac{t-b}{a} \right) dt .$$

На рис. 4 приведена скейлограмма - результат непрерывного вейвлет-анализа (вейвлет Гаусса) рассмотренного на рис 1 временного ряда.

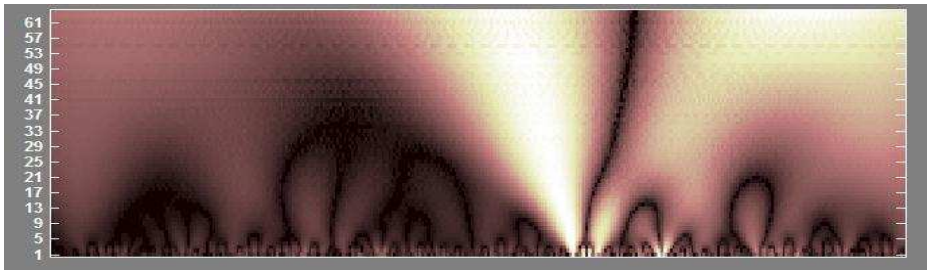

Рис. 4. Скейлограмма временного ряда (вейвлет Гаусса), приведенного на рис. 1

Предложенный метод визуализации абсолютных отклонений ΔL, как и метод вейвлет-преобразований, позволяет выявлять единичные и нерегулярные «всплески», резкие изменения значений ряда в разные периоды времени. Следует отметить, что метод вейвлет-преобразований может применяться с использованием разнообразных вейвлетов. Применение другого вейвлета Хаара (рис. 5) также не позволило идентифицировать ряд особенностей исходного ряда измерений в ноябре 2008 года, по меньшей мере, эти особенности не показаны как скелетоны на рис. 5 *б)*. Метод ΔL, реализация которого существенно проще, позволил определить эту аномалию.

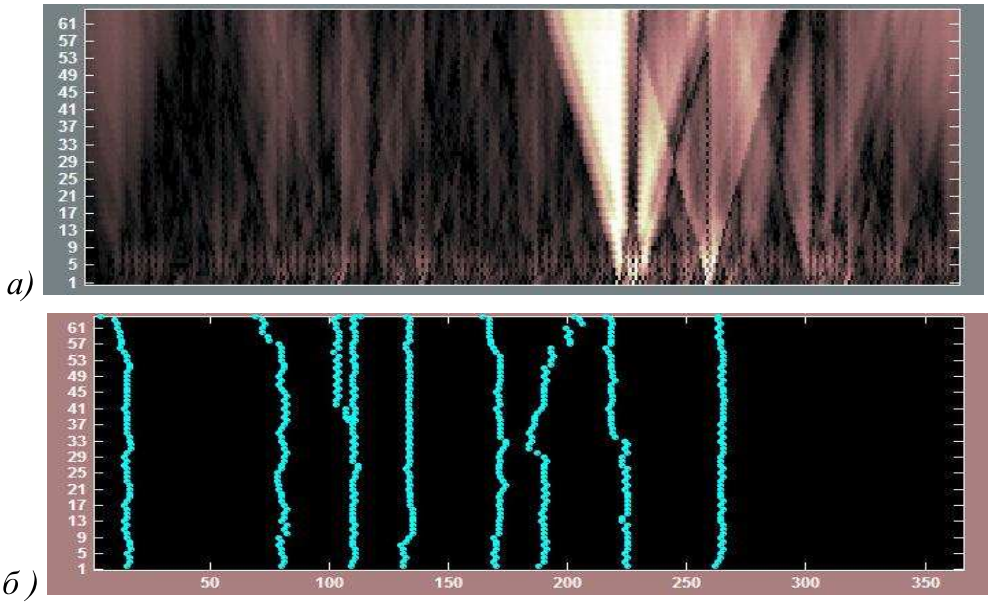

Рис. 5. Скейлограмма временного ряда (вейвлет Хаара), приведенного на рис. 1: а) скейлограмма (ось абсцисс – день года, ось ординат - частота); б) линии локальных максимумов скейлограммы

Выявление гармоник

ΔL -метод оказывается достаточно эффективным для выявления гармонических составляющих исследуемого ряда. На рис. 6 показана ΔL -диаграмма ряда, соответствующего синусоиде ($y(i) = \sin(i\pi / 7),\ \ i = 1,...,366$). Применение ΔL -метода к ряду, составленному из количества публикаций, сосканированных системой InfoStream из Интернет без учета тематического деления, имеет явно выраженную гармоническую составляющую (общее количество публикаций зависит от дня недели), что можно видеть на рис. 7. Кроме того, на этой диаграмме заметны отклонения от общей динамики объемов публикаций в периоды праздничных дней.

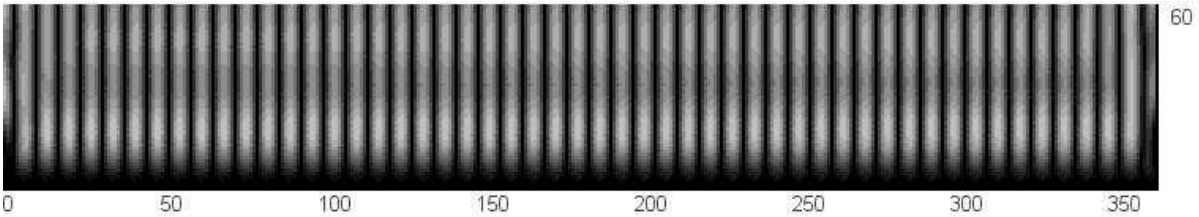

Рис. 6. ΔL -диаграмма синусоиды

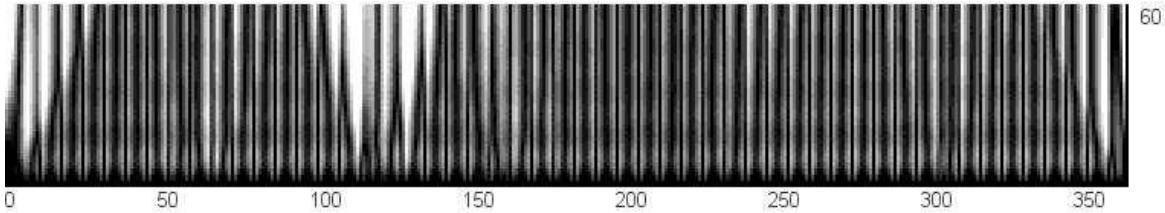

Рис. 7. ΔL -диаграмма ряда из количества публикаций, сосканированных ежесуточно системой InfoStream в 2008 году

Анализ интернет-трафика

В результате исследований, описанных в [7], известно, что изначально не проявляющие свойств самоподобия потоки данных в сети Интернет, пройдя обработку на узловых серверах, приобретают фрактальные свойства. Однако, авторами показано, что автокорреляционные свойства трафика, проходящего между крупными провайдерами, не обладает явно выраженными свойствами самоподобия, что подтверждается поведением показателя $F(L)$, т.е. метод DFA не позволяет делать какие-либо выводы о характере процесса. Вместе с тем, отдельные, не усредненные отклонения указанного трафика дают достаточно полную и наглядную картину в различные интервалы времени – именно это подтверждается методом ΔL. На рис. 8 представлена посуточная диаграмма трафика, проходящего между двумя провайдерами в течение 2008 года.

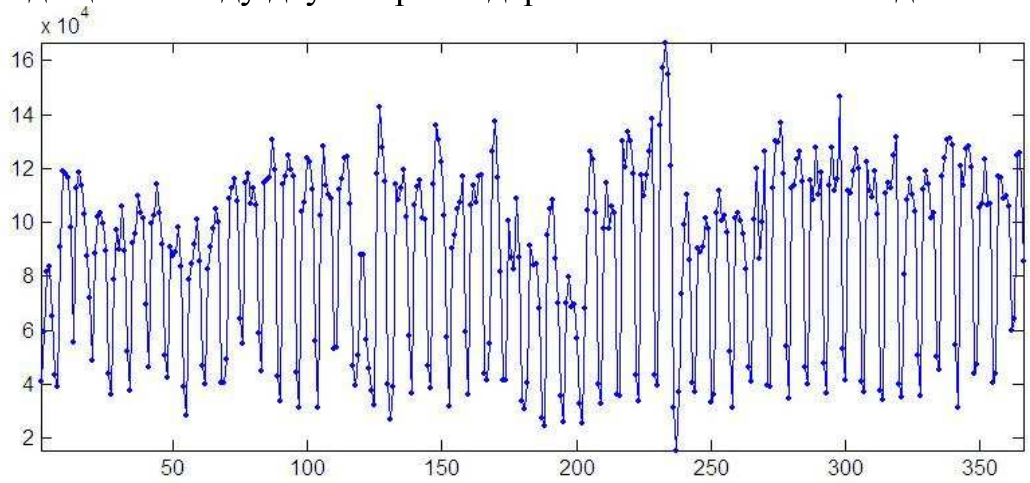

Рис. 8. Посуточная диаграмма трафика между провайдерами Интернет

Даже опытный эксперт без специального инструментария не может определить зоны регулярных отклонений посуточного трафика от общей тенденции. В этом случае применение ΔL-метода, позволяет выявить временные периоды, в течение которых процесс имеет специфический характер (рис. 9), что можно учитывать при планировании нагрузки телекоммуникационных узлов. На приведенной ΔL-диаграмме видно,

что такими специфическими периодами оказался летний сезон и начало осени (первые проявления кризисных явлений в экономике).

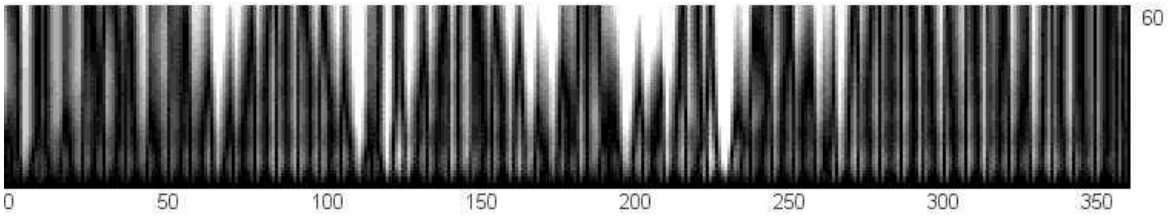

Рис. 9. ΔL -диаграмма диаграмма трафика между провайдерами

Выводы

В случае применения ΔL-метода, который достаточно прост в программной реализации, не требуется решать сложную задачу выбора и обоснования применения подходящего вейвлета; в отличие от методов фрактального анализа предложенный подход не требует значительных объемов точек ряда измерений.

Литература

1. Peng C.K. Mosaic organization of DNA nucleotides / Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL. // Phys Rev E. -1994, 49 (2) 1685-1689.
2. Peng C.K. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series / Peng C.K, Havlin S., Stanley H.E., Goldberger A.L. // Chaos. - Vol. 5. - 1995. - pp. 82.
3. Ландэ Д.В. Сканер системы контент-мониторинга InfoStream // Открытые инф. и компьютерные интегрированные технологии: Сб. науч. трудов. Вып. 28. –Х.: "ХАИ", 2005. - С. 53-58.
4. Додонов А.Г. Самоподобие массивов сетевых публикаций по компьютерной вирусологии / Додонов А.Г., Ландэ Д.В. // Реєстрація, зберігання і обробка даних, -2007. - Т. 9, - N 2. - С. 53-60.
5. Астафьева Н.М. Вейвлет-анализ: основы теории и примеры применения // УФН. -1996. - Т. 166. - № 11. – С. 1145-1170.
6. Buckheit J. Wavelab and reproducible research / Buckheit J., Donoho D. // Stanford University Technical Report 474: Wavelets and Statistics Lecture Notes, 1995. -27 p.
7. Feng W. The failure of TCP in high-performance computational grids / Feng W., Tinnakornsrisuphap P. // Proceedings of the 2000 ACM/IEEE conference, p. 37.