УДК 004.67

А. А. Снарский[1], Д. В. Ландэ[2]
[1]НТУУ «Киевский политехнический институт»
Проспект Победы, 37, 03056 Киев, Украина
[2]Институт проблем регистрации информации НАН Украины
ул. Н. Шпака, 2, 03113 Киев, Украина

Графы видимости — инструмент сетевого анализа рядов измерений

Приведен обзор методов сетевого анализа рядов измерений, базирующихся на алгоритмах построения графов видимости. Описаны оригинальные алгоритмы построения динамического графа видимости и компактифицированного графа горизонтальной видимости для сети слов.

Ключевые слова: *граф видимости, динамический граф видимости, сложная сеть, сеть слов, ряд измерений, цифровая обработка сигналов.*

Анализ временных рядов играет важную роль во многих областях науки и техники — в физике, биологии, сейсмологии, лингвистике, экономике и т.д., в частности, при определении скрытой периодичности, при решении задач диагностирования и прогноза.

Анализ сложных временных рядов сегодня базируется на использовании многочисленных методов — статистических, корреляционных, фрактальных. Например, вычисление константы Херста [1] позволяет определить персистентность ряда, возможность следования своему предыдущему тренду (т.е., если значения ряда возрастали, то они будут возрастать и дальше, а если падали, то падение продолжится). В [2] был предложен дисперсионный метод фрактального анализа DFA (Detrended Fluctuation Analysis), успешно применяемый при анализе распознавания сердечнососудистых заболеваний.

В последние несколько лет появилось новое направление в исследовании временных рядов с нетривиальной структурой, использующее хорошо развитые методы анализа сложных сетей, базирующееся на так называемых «графах видимости». Временным рядам при этом по определенному алгоритму ставятся в соответствие такие графы, свойства которых (распределение узлов по степеням, кластеризация, ассортативность и т.п.) активно изучаются в настоящее время. И если, например, метод расчета константы Херста в настоящее время хорошо изучен — при $H > 1/2$ направление динамики исследуемого ряда сохраняется, в про-

тивном случае — изменяется, то найдя ту или иную сетевую характеристику графа видимости, часто заранее не очевидно, какую информацию об исходном ряде она несет. Необходимо отметить, что временной ряд со сложной структурой содержит (или может содержать) большой набор характеристик процесса его порождающего. Поэтому любые новые, в том числе сетевые характеристики, могут оказаться полезными.

Существует несколько алгоритмов отображения временного ряда в сложную сеть, например в [3], было предложено в качестве веса ребер графа использовать близость координат в сечении Пуанкаре исходного временного ряда [4–8]. В другом типе алгоритмов вводится так называемый «граф видимости». В работе [9] был предложен алгоритм построения графа взаимной видимости (Natural Visibility Graph, NVG-algorithm). Несколько позже был предложен другой алгоритм, подобный NVG — «граф горизонтальной видимости» (Horizontal Visibility Graph, HVG) [10].

Алгоритм построения графов видимости проиллюстрирован на рис. 1. При построении NVG-графа на горизонтальной оси (ось времени) отмечаются точки t_i, от которых в перпендикулярном направлении строятся отрезки высотой, равной значениям ряда измерений в этих точках — $x(t_i)$. Узлами NVG-графа являются внешние вершины построенных отрезков. Связь между вершинами в NVG-графе считается существующей, если прямая, соединяющая соответствующие вершины отрезков, не пересекает ни одного из построенных отрезков, находящихся между. В алгоритме HVG вертикальные отрезки соединяются горизонтально, подробное описание приведено ниже.

На рис. 1 вверху (слева) схематически изображен критерий видимости для угла зрения $\alpha = \pi / 2$. Связи, отвечающие углам, меньшим угла зрения (например α_1), обозначены толстыми линиями, а отвечающие углам, большим угла зрения (например α_2) — тонкими линиями. Оба этих типов линий (и тонкие и жирные) образуют натуральный граф видимости (NVG). Графу динамической видимости (DVG) отвечают только связи, обозначенные жирными линиями, наклон которых меньше угла зрения. В нижней части на рис. 1 изображен процесс построения HVG.

Алгоритмы NVG и HVG были использованы при исследовании временных рядов сложной структуры, связанных с самыми различными явлениями: пульсацией турбулентных течений [11], индексами фондового рынка [12], сердцебиениями человека [13, 14], при изучении стохастических и хаотических временных рядов и для многих других приложений.

Как в NVG, так и HVG — каждому временному ряду соответствует свой граф.

Известно, что граф взаимной видимости NVG обладает следующими свойствами: 1) все узлы графа, соответствующие значениям временного ряда, являются смежными для узлов, соответствующих «соседним» значениям исходного ряда; 2) связи являются ненаправленными (хотя возможно обобщение и на направленные связи); 3) граф видимости инвариантен относительно афинных преобразований; 4) отображение временного ряда в NVG является алгоритмом с потерями, например, два разных временных ряда {3, 10, 3, 10, …} и {2, 15, 2, 15, …} приводят к одному и тому же NVG.

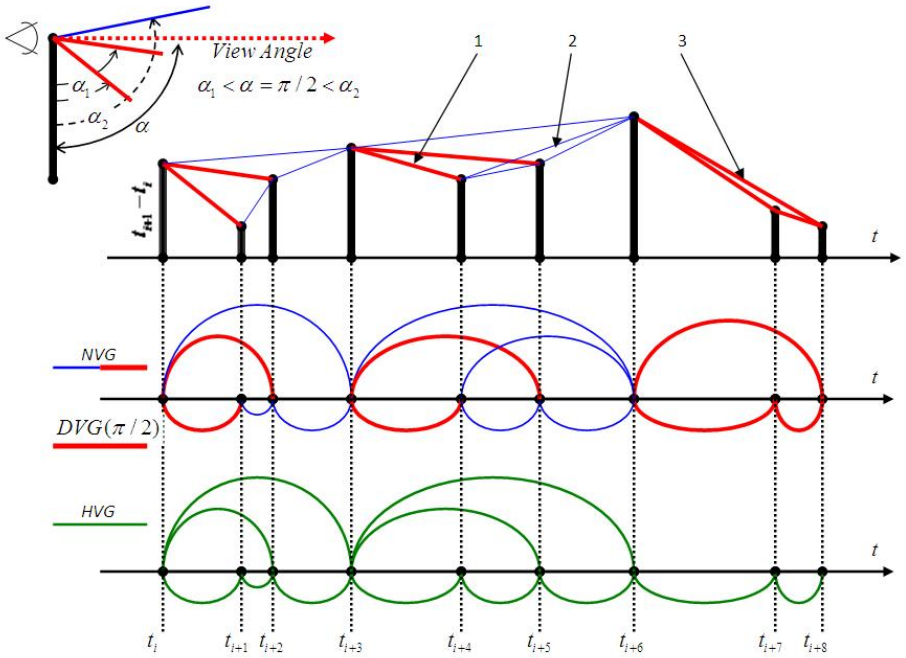

Рис. 1. Схема построения графов видимости

В работе [10] был предложен алгоритм построения графа горизонтальной видимости — HVG.

Между узлами, соответствующими элементам временного ряда, существует связь, если они находятся в «прямой видимости», т.е. если их можно соединить горизонтальной линией, не пересекающей никакую другую вертикальную линию. Этот (геометрический) критерий можно записать, согласно [15, 16] следующим образом: два узла (элемента ряда), например I^n и I^m, соединены связью, если (см. рис. 2) $\sigma_n, \sigma_m > \sigma_p$ для всех $n < p < m$.

Алгоритм построения можно представить удобным для вычисления способом. Так, например, на рис. 2 для узла I^{n+2} смежными в сети считаются слова I^n и I^{n+5} (и устанавливаются ребра-связи), такие что I^n — ближайший слева от I^{n+2} элемент, со значением σ_n, превышающем оценку элемента I^{n+2}, а $I^m\left(m = n+5\right)$ — ближайший справа от I^{n+2} элемент, для которого $\sigma_m > \sigma_{n+2}$.

Граф горизонтальной видимости эффективно применяется для выявления скрытых периодичностей во временных рядах. В частности, для периодических временных рядов HVG приводит к следующему соотношению для средней степени узлов $\bar{K}\left(T\right)$ соответствующего графа

$$\bar{K}\left(T\right) = 4\left(1 - \frac{1}{2T}\right), \tag{1}$$

где T — период временного ряда.

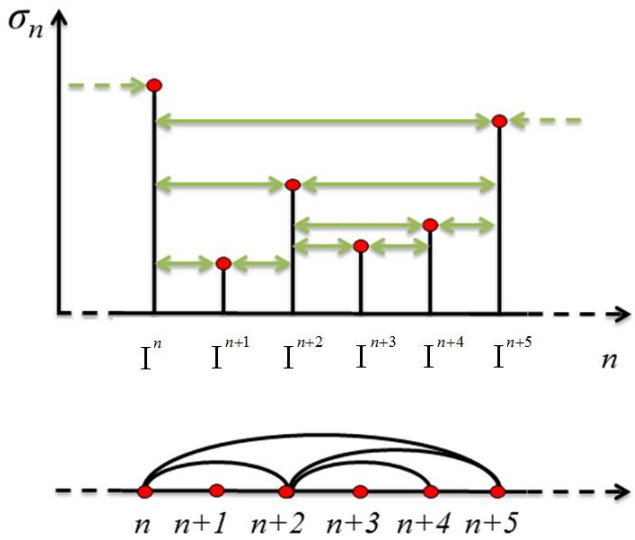

Рис. 2. Пример построения графа горизонтальной видимости

Для непериодического временного ряда ($T \to \infty$) при этом получается $\bar{K}(T) = 4$.

На базе соотношения (1) может быть построен фильтр для определения скрытых периодичностей во временных рядах.

Распределение степеней узлов HVG — $P(k)$ для случайного некоррелированного временного ряда является экспоненциальным [10]:

$$p(k) = \frac{1}{2}\left(\frac{2}{3}\right)^{k-2}, \quad k = 2,3,... \tag{2}$$

В работе [15] было предложено обобщение NVG-алгоритма, названное алгоритмом графа динамической видимости (DVG — Dynamical Visibility Graph) — рис. 1. Дополнительным параметром (к алгоритмам NVG и HVG) алгоритма построения DVG является «угол зрения» α. Для каждой из связей стандартного графа взаимной видимости можно вычислить угол наклона по отношению к горизонтальной оси. Связями графа динамической видимости будут только те, наклон которых менее заданного угла — «угла зрения» α.

Угол зрения α — непрерывно изменяющийся параметр, каждому значению этого параметра соответствует свой граф, являющийся подграфом стандартного графа взаимной видимости (NVG). Построенный алгоритм является обобщением алгоритма NVG, при $\alpha = \pi$ DVG-граф переходит в NVG-граф. Предложенный DVG-алгоритм позволяет исследовать зависимость параметров графа от угла зрения α (форма зависимости, скорости ее роста, скачки и др.). Возможность изменять произвольно угол зрения добавило в название алгоритма слово динамический, далее будем использовать также обозначение DVG (α).

Построение DVG-графа происходит следующим образом (рис. 1). Временной ряд представляется как последовательный набор моментов времени

$\{t_i,\ i=1,...,N\}$, в которые происходят некоторые события, например R — пик в кардиограмме. Вначале, сопоставим этому набору временной ряд $\{x(t_i)=t_{i+1}-t_i,\ \ i=1,...,N-1\}$. На плоскости на горизонтальной оси отмечаются точки t_i, от которых в перпендикулярном направлении строятся отрезки высотой $x(t_i)$. Узлами графа взаимной видимости являются внешние вершины построенных отрезков. Связь между вершинами считается существующей, если прямая, соединяющая соответствующие вершины отрезков не пересекает ни одного из построенных отрезков, находящихся между ними. Для каждой из связей стандартного графа взаимной видимости можно вычислить угол наклона по отношению к горизонтальной оси. Связями графа динамической видимости будут только те связи стандартного графа, наклон которых менее заданного угла — «угла зрения» α.

Формальный критерий видимости, т.е. условие, при выполнении которого связь NVG-графа будет принадлежать DVG (α) можно записать следующим образом. Рассмотрим два произвольных момента времени t_i и t_k, $(i<k)$ и все моменты времени между ними t_j, $i<j<k$. Для DVG (α) критерий видимости, т.е. существования связи между узлами *i* и *k* в NVG:

$$x_k < x_i + \left(x_j - x_i\right)\frac{t_k - t_i}{t_j - t_i},\quad i<j<k, \tag{3}$$

должен быть дополнен условием, ограничивающим угол зрения α:

$$\alpha > \alpha_{ik} = \operatorname{arctg}\frac{x_k - x_i}{t_k - t_i}. \tag{4}$$

Рассмотрим теперь одну из характеристик DVG-графа — относительное количество кластеров — $Q(\alpha)$, где под кластером мы будем понимать совокупность связанных между собой узлов, не связанных с другими кластерами. Таким образом, $Q(\alpha)$ — число кластеров деленное на полное число узлов графа. При этом единичный узел, не связанный с другими узлами, кластером не считается.

Совершенно очевидно, что при малом угле зрения связей между узлами нет, и число кластеров равно числу узлов, т.е. максимально. При этом $Q(\alpha<\pi/4)=1$. При увеличении угла зрения (начиная с $\alpha=\pi/4$) начинают появляться связи между узлами, размер кластеров растет, а их число монотонно падает. При достижении критического значения угла $\alpha_c=\pi/2$ практически все узлы соединены между собой, граф состоит из нескольких кластеров и относительное число кластеров порядка $1/N \ll 1$. Как показало численное моделирование, относительное число кластеров $Q(\alpha)$ для достаточно длинных временных рядов ведет себя вблизи критического значения α_c степенным образом (аналогично параметру порядка в теории фазовых переходов второго рода):

$$Q(\alpha) \sim (\alpha_c - \alpha)^{\beta}. \tag{5}$$

Значение критического показателя β является характеристикой временного ряда. Как показали расчеты, его значения являются разными для разных временных рядов. Ниже приведен пример применения DVG(α) алгоритма к анализу экспериментальных данных — RR-интервалов кардиограмм, взятых из базы данных PhysioNet [16].

Было взято 72 серии RR-интервалов здоровых людей, 44 серии с Congestive Heart Failure (Застойная сердечная недостаточность) и 25 серий с аритмией. Длина временного каждого ряда была разной, между 6 и 10 на 10^4 интервалов. Для каждого из типов интервалов была построена и усреднена зависимость $Q(\alpha)$. На рис. 3 показаны эти зависимости. Как видно из рисунка, существуют такие диапазоны углов, на которых разные типы интервалов четко различаются.

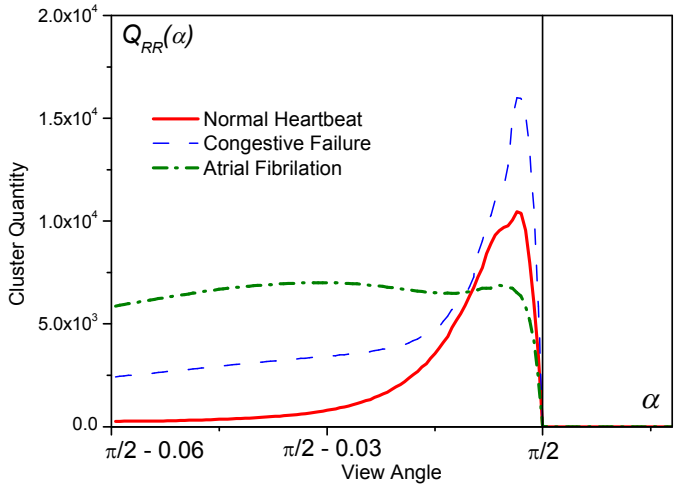

Рис. 3. Угловые зависимости $Q(\alpha)$ для разных типов RR-интервалов

Построение сетей слов, узлами которых являются элементы текста, позволяет выявлять структурные элементы, без которых этот текст теряет свою связность. При этом актуальной является задача определения того, какие из важных структурных элементов оказываются также информационно-значимыми, определяющими информационную структуру текста. Такие элементы могут использоваться также для идентификации еще не достаточно четко определенных компонент текста, таких как коллокации, сверхфразовые единства [17], например, при поиске подобных фрагментов в различных текстах [18].

Известно несколько подходов к построению сетей из текстов, так называемых сетей слов (Language Network), и различные способы интерпретации узлов и связей, что приводит, соответственно, к различным видам представления таких сетей. Узлы могут быть соединены между собой, если соответствующие им слова стоят рядом в тексте [19, 20], принадлежат одному предложению или абзацу [21], соединены синтаксически [22, 23] или семантически [24, 25].

Подходы, связанные с графами видимости, позволяют строить сетевые структуры на основании текстов, в которых отдельным словам или словосочетаниям некоторым специальным образом поставлены в соответствие числовые весовые значения. В качестве функции, ставящей в соответствие слову число, можно рассматривать, например, порядковый номер уникального слова в тексте, длину слова, «вес» слов в текстах, общепринятую оценку TFIDF (в каноническом виде, равную произведению частоты слова в фрагменте текста — term frequency — на двоичный логарифм от величины, обратной количеству фрагментов текста, в которых это слово встретилось — inverse document frequency) или ее варианты [26, 27], а также другие весовые оценки.

В качестве весовой оценки TFIDF из полного текста, состоящего из N слов, текст разбивается на фрагменты, содержащие заданное количество слов M (например $M = 500$). Затем для каждого слова i, входящего в текст, подсчитывается количество фрагментов $df(i)$, в которые это слово входит, а также общее количество вхождений данного слова i в текст — $n(i)$. После этого по формуле

$$tfidf(i) = \frac{n(i)}{N} \log\left(\frac{N}{M \times df(i)}\right) \tag{6}$$

рассчитывается среднее значение TFIDF весовой оценки каждого слова.

При построении сетей слов также может быть использована дисперсионная оценка важности слов [28], которая реализуется следующим образом: пусть текст состоит из N слов ($n = 1,...,N$, n — порядковый номер слова в тексте, позиция слова). Некоторое слово, например A, обозначается как A_k^n, где индекс $k = 1,2,...,K$ — номер появления данного слова в тесте, а n — позиция данного слова в тексте. Например A_3^{50}, означает, что на 50-й позиции текста находится слово A, которое встретилось третий раз.

Интервал между последовательными появлениями слова при таких обозначениях будет величина $\Delta A_k = A_{k+1}^m - A_k^n = m - n$, где на m-й и n-й позициях в тексте находится слово A, которое встретилось $k + 1$ и k-й разы.

Предложенная в [28] дисперсионная оценка рассчитывается как

$$\sigma_A = \frac{\sqrt{\left\langle \Delta A^2 \right\rangle - \left\langle \Delta A \right\rangle^2}}{\left\langle \Delta A \right\rangle}, \tag{7}$$

где $\left\langle \Delta A \right\rangle$ — среднее значение последовательности $\Delta A_1, \Delta A_2, ..., \Delta A_K$; $\left\langle \Delta A^2 \right\rangle$ — последовательности $\Delta A_1^2, \Delta A_2^2, ..., \Delta A_K^2$; K — количество появления слова A в тексте.

В отличие от остальных рядов, изучаемых в рамках цифровой обработки сигналов, ряды из цифровых значений, соответствующих словам, преобразуются в графы горизонтальной видимости, в которых узлам соответствуют не только цифровые значения, но сами слова, выражающие определенное смысловое значение.

Сеть слов с использованием алгоритма горизонтальной видимости строится в три этапа. На первом на горизонтальной оси отмечается ряд узлов, каждый из которых соответствует словам в порядке появления в тексте, а по вертикальной оси откладываются весовые численные оценки (визуально – набор вертикальных линий, см. рис. 1).

На втором этапе по алгоритму, описанному выше, строится граф горизонтальной видимости.

На третьем, заключительном этапе, полученная на предыдущем этапе сеть компактифицируется. Все узлы с данным словом, например словом *A*, объединяются в один узел (естественно, индекс и номер положения слова при этом исчезают). Все связи таких узлов также объединяются. Важно отметить, что между любыми двумя узлами при этом остается не более одной связи — кратные связи изымаются. В частности это означает, что степень (число связей) узла *A* не превышает суммы степеней $\sum_k A_k^n$. В результате получается новая сеть слов — *компактифицированный граф горизонтальной видимости* (CHVG) — рис. 4.

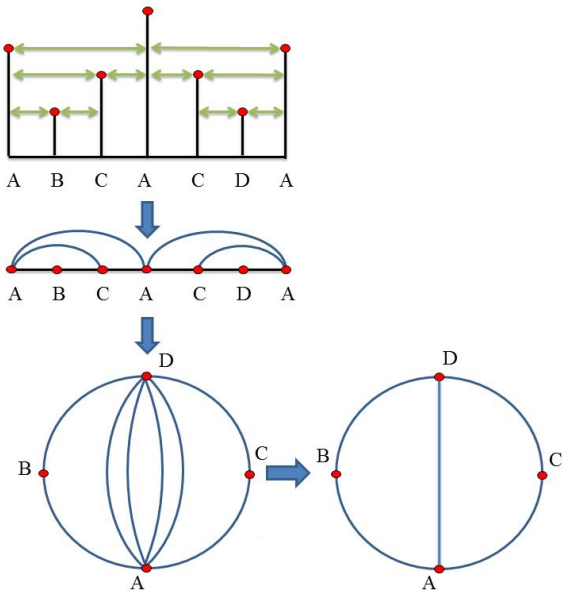

Рис. 4. Этапы построения компактификационного графа горизонтальной видимости

Для всех построенных CHVG-сетей слов было определено распределение степеней узлов, которое оказалось близким к степенному ($p(k) = Ck^{\alpha}$), т.е. эти сети являются безмасштабными. Были проведены расчеты параметров сетей для всех рассмотренных литературных произведений. В результате оказалось, что для всех из них коэффициент α изменялся в диапазоне от –1 до –0,97.

В состав узлов с наибольшими степенями для CHVG-сетей, наряду с личными местоимениями и другими служебными словами (частицы, предлоги, союзы и т.д.), попадают слова, определяющие информационную структуру текста [29, 30].

В результате проведенных исследований получены следующие результаты.

1. Рассмотрены различные алгоритмы построения графов видимости.

2. Исследован предложенный алгоритм динамического графа видимости (DVG).

3. В угловых зависимостях параметров DVG найдены аналоги фазового перехода второго рода и вычислен критический индекс.

4. Показано, что анализ RR-интервалов, проведенный методом DVG, позволяет разделять на классы различные отклонения в работе сердечнососудистой системы.

5. Предложен оригинальный алгоритм построения компактифицированного графа горизонтальной видимости (CHVG) для сетей слов.

6. На основе последовательностей дисперсионных оценок слов текста и оценок IFIDF построены CHVG-сети слов для различных текстов.

7. Для литературных текстов среди узлов, соответствующих CHVG, с наибольшими степенями присутствуют слова, не только обеспечивающие связность структуры текста, но и определяющие его информационную структуру, отражают семантику литературных произведений.

В целом, можно утверждать, что анализ сложных временних рядов методами сложных сетей являетя полезным и перспективным методом.

1. *Федер Е.* Фракталы / Е. Федер. — М: Мир, 1991. — 254 с.

2. *Mosaic* Organization of DNA Nucleotides / C.-K. Peng, S.V.Buldyrev, S. Havlin [et al.] // Phys.Rev. — 1994. — **E 49**. — P. 1685–1689.

3. *Small M.* Applied Nonlinear Time Series Analysis: Applications in Physics, Physiology and Finance / M. Small // World Scient. Publ. — 2005. **— 52**. — 246 p.

4. *Zhang J.* Complex Network From Pseudoperiodic Time Series: Topology Versus Dynamics / J. Zhang, M. Small // Phys. Rev. Lett. — 2006. **— 96**. — P. 238701.

5. *Zang J.* Detecting Chaos in Pseudoperiodic Time Series Without Embedding / J. Zang, X. Luo, M. Small // Phys. Rev. — 2006. — **E 73**. — P. 016216.

6. *Characterizing* Pseudoperiodic Time Series Through Complex Network Approach / J. Zhang, J. Sun, X. Luo [et al.] // Physica D. — 2008. **— 237**. — P. 2856–2865.

7. *Xu X.* Superfamily Phenomena and Motifs of Networks Induced from Time Series / X. Xu, J. Zhang, M. Small // Proc. Natl. Acad. Sci. U.S.A. — 2008. — N 105(50). — P. 19601–19605.

8. *Ambiguities* in Recurrence-Based Complex Network Representations of Time Series / R.V. Donner, Y. Zou, J.F. Donges [et al.] // Phys. Rev. — 2010. — **E 81**. — P. 015101.

9. *From* Time Series to Complex Networks: the Visibility Graph / L. Lacasa, B. Luque, F. Ballesteros [et al.] // Proc. Natl. Acad. Sci. U.S.A. — 2008. — N 105. — P. 4972.

10. *Horizontal* Visibility Graphs: Exact Results for Random Time Series / Luque B., Lacasa L., Ballesteros F., Luque J. // Phys. Rev. — 2009. — **E 80**. — P. 046103.

11. *Liu C.* Statistical Properties of Visibility Graph of Energy Dissipation Rates in Three-Dimensional Fully Developed Turbulence / C. Liu, W.-X. Zhou, W.-K. Yaun // Physica A. — 2010. — N 389(13). — P. 2675–2681.

12. *Qian M.-C.* Universal and Nonuniversal Allometric Scaling Behaviours in the Visibility Graphs of World Stock Market Indices / M.-C. Qian, Z.-Q. Jiang, W.-X. Zhou // J. Phys. A: Math. Their. — 2010. **—** N 43(33). — P. 335002.

13. *Shao Z.-G.* Network Analysis of Human Heartbeat Dynamics / Z.-G. Shao // Appl. Phys. Lett.— 2010. — **96**. — P. 073703.

14. *Li X.* Detection and Prediction of the Onset of Human Ventricular Based on Complex Network Theory / X. Li, Z. Dong // Phys. Rev. — 2011. — **E 84**. — P. 062901.

15. *Bezsudnov I.V.* From Time Series to Complex Networks: the Dynamical Visibility Graph / I.V. Bezsudnov, S.V. Gavrilov, A.A. Snarskii // arXiv:1208.6365. — 2012. — 13 p.

16. *Moody G.B.* Predicting Acute Hypotensive Episodes: The 10th Annual PhysioNet / G.B. Moody, L.H. Lehman // Computers in Cardiology Challenge. — 2009. — N 36. — P. 541–544.

17. *Солганик Г.Я.* Синтаксическая стилистика. Сложное синтаксическое целое / Г.Я. Солганик. — [2-е изд., испр. и доп.]. — М.: Высш. школа, 1991. — 182 с.

18. *Broder A.* Identifying and Filtering Near-Duplicate Documents, COM'00 / Broder A. // Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching, 2000. — P. 1–10.

19. *Ferrer-i-Cancho R.* The Small World of Human Language / R. Ferrer-i-Cancho, R.V. Sole // Proc. R. Soc. Lond. — 2001. — **B 268**. — P. 2261.

20. *Dorogovtsev S.N.* Language as an Evolving Word Web / S.N. Dorogovtsev, J.F.F. Mendes // Proc. R. Soc. Lond. — 2001. — **B 268**. — P. 2603.

21. *The Network* of Concepts in Written Texts / S.M.G. Caldeira, T.C. Petit Lobao, R.F.S.Andrade [et al.] // Preprint Arxiv, 2005. — physics/0508066.

22. *Ferrer-i-Cancho R.* Patterns in Syntactic Dependency Networks / R. Ferrer-i-Cancho, R.V. Sole, R. Kohler // Phys. Rev. — 2004. — **E 69**. — P. 051915.

23. *Ferrer-i-Cancho R.* The Variation of Zipf's Law in Human Language / R. Ferrer-i-Cancho // Phys. Rev. — 2005. — **E 70**. — P. 056135.

24. *Topology* of the Conceptual Network of Language / A.E. Motter, A.P S. de Moura, Y.-C. Lai, P. Dasgupta // Phys. Rev. — 2002. — **E 65**. — P. 065102(R).

25. *Sigman M.* Global Properties of the Wordnet Lexicon / M. Sigman, G.A. Cecchi // Proc. Natl. Acad. Sci. USA. — 2002. — **99**. — P. 1742.

26. *Jones K.S.* A Statistical Interpretation of Term Specificity and Its Application in Retrieval / K.S. Jones // Journal of Documentation. — 1972. — **28**(1). — P. 11–21.

27. *Salton G.* Introduction to Modern Information Retrieval / G. Salton, M. J. McGill. — New York: McGraw-Hill, 1983. — 448 p.

28. *Keyword* Detection in Natural Languages and DNA / M. Ortuño, P. Carpena, P. Bernaola [et al.] // Europhys. Lett. — 2002. — **57**(5). — P. 759–764.

29. *Черняховская Л.А.* Смысловая структура текста и ее единицы / Л.А. Черняховская // Вопросы языкознания. — 1983. — № 6. — С. 118–126.

30. *Giora R.* Segmentation and Segment Cohesion: On the Thematic Organization of the Text / R. Giora // An Interdisciplinary Journal for the Study of Discourse Amsterdam. — 1983. — **3**, N 2. — P. 155–181.

Поступила в редакцию 03.06.2013