УДК 004.067

А.А. Снарский, Д.В. Ландэ

Графы видимости – инструмент исследования сложных рядов

Приведен обзор методов построения графов видимости для сетевого анализа временных рядов. Приведено описание оригинального алгоритма построения графа горизонтальной видимости для текстов – сети слов.

Ключевые слова: *граф горизонтальной видимости, сложные сети, сеть слов, временной ряд, цифровая обработка сигналов*

Изучение рядов со сложной, в том числе фрактальной, структурой – актуальная задача. В значительной мере это связано с большим прикладным потенциалом таких исследований, которые охватывают огромный диапазон явлений от сердечного ритма до землетрясений [1,2]. При исследовании временных рядов используются различные методы: обычная статистика (вычисление средних значений и различных моментов), спектры мощности (например, анализ наличия 1/f шума), определение фрактальных характеристик, некоторые из которых свидетельствуют о достаточно нетривиальных свойствах этих рядов, например, о мультифрактальности, существовании странных аттракторов и т. п. Анализ временных рядов методом построения графов динамической видимости [3], в частности, позволяет рассчитать характеристику, ведущую себя аналогично параметру порядка в теории фазовых переходов второго рода.

В последнее время появился и активно развивается метод исследования временных рядов, в основе которого лежит их преобразование в граф (сложную сеть). При таком отображении объединяются две развитые области исследований – нелинейные методы анализа временных рядов и методы теории сложных сетей [4-6]. Появляется возможность применить богатые, хорошо развитые методы анализа сложных сетей к анализу сложных по структуре временных рядов.

Существует несколько типов алгоритмов отображения временного ряда в сложную сеть. Первый тип [7] использует близость координат в сечении Пуанкаре временного ряда. Другой тип алгоритма вводит понятие т.н. «графа видимости». В работе [8] был предложен алгоритм построения графа видимости (Natural Visibility Graph, алгоритм NVG). Несколько позже был предложен «близкий по духу» к NVG алгоритм построения «горизонтального графа видимости» (Horizontal Visibility Graph, HVG-алгоритм) [9]. Как в NVG так и HVG – каждому временному ряду соответствует свой граф. Применение алгоритмов NVG и HVG позволило описать и исследовать временные ряды сложной структуры, связанные с самыми различными явлениями: пульсациями турбулентных течений,

индексами фондовых рынков, сердечными ритмами, стохастическими и хаотическими временными рядами и многими другими.

В работе [3] было предложено обобщение алгоритма NVG – алгоритм построения графа динамической видимости (Dynamical Visibility Graph, DVG-алгоритм). В этом случае каждой связи NVG-графа ставится в соответствие величина, называемая «угол зрения». Связями графа динамической видимости будут только те связи NVG-графа, «угол зрения» которых менее заданного угла – «угла зрения» α, таким образом, для каждого "угла зрения" строится свой DVG-граф. Алгоритм DVG 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».

Построение сетей, узлами которых являются элементы текста, слова или словосочетания, фрагменты естественного языка, в некоторых случаях позволяет выявлять структурные элементы текста, без которых он теряет свою связную структуру, информационно-значимые элементы, а также второстепенные для понимания текста слова. Такие сети могут использоваться также для идентификации таких нетрадиционных компонент текста, как колокации, сверхфразовые единства [10], для нахождения подобных фрагментов разных текстов [11]. Существует множество подходов к построению сетей из текстов, так называемых сетей слов (language network), и различные способы интерпретации узлов и связей, что приводит, соответственно, к различным представлениям таких сетей. Узлы могут быть соединены между собой, если соответствующие им слова стоят рядом в тексте [12], принадлежат одному предложению [13], соединены синтаксически [14] или семантически [15].

Базируясь на подходах графов видимости – NVG, HVG, DVG, из текстов, в которых отдельным словам или словосочетаниям некоторым образом поставлены в соответствие числовые значения, также можно строить сетевые структуры. В качестве функции, ставящей в соответствие слову число, можно, например, рассматривать порядковый номер уникального слова в тексте, длину слова, «веса» слов в текстах, например, общепринятую оценку TFIDF, другие весовые оценки слова в тексте.

При этом, ряды полученные из текстов (каждому слову ставится в соответствие цифра, например, порядковый номер этого слова в словаре языка) отличаются от «обычных» рядов тем, что значительное число элементов этого ряда имеют тождественные численные значения – число таких значений равно числу уникальных слов в тексте и оно намного меньше числа элементов рядов.

Предлагаемая авторами сеть слов с использованием алгоритма горизонтальной видимости строится в два этапа [16]. На первом этапе строится традиционный граф горизонтальной видимости. Для этого на горизонтальной оси отмечается ряд узлов, каждый из которых соответствует

словам в порядке появления в тексте, а по вертикальной оси откладываются весовая оценка слова (визуально – набор вертикальных линий, см. рис.1). Между узлами 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 следующим образом: два узла (слова) слова, например, $B_3(n)$ и $C_7(m=n+5)$ соединены связью, если, см. рис. 1, $\sigma_n, \sigma_m > \sigma_p$, для всех $n < p < m$.

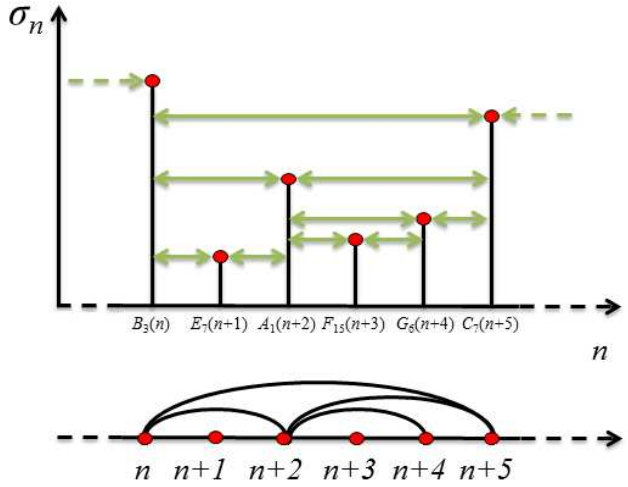

Рис. 1. Пример построения графа горизонтальной видимости

На втором этапе, полученная на первом 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 узлами при этом существует 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 горизонтальный граф видимости* (КГГВ).

В качестве текстов при построении КГГВ рассматривались роман романа М. Булгакова «Мастер и Маргарита», роман Г.Мелвилла «Моби Дик» и массивы новостной информации из веб-пространства.

Для таких сетей, построенных на основании текстов литературных произведений и потоков новостей, выявлены феномены «клуба богатых». Выявлен факт степенного распределения степеней вершин построенных КГГВ-сетей. В частности, в КГГВ-сети, соответствующей роману М. Булгакова «Мастер и Маргарита» в состав узлов с наибольшими степенями попали такие слова, как Иван, Мастер, Варенуха, Берлиоз, Бегемот, Римский, профессор, Левий, Иешуа. В состав узлов с наибольшими степенями по роману Г. Мелвилла «Моби Дик, или Белый кит» попали слова: whale, Ahab, boat, Queequeg, sperm, Starbuck, Pequod.

В результате исследований сетей слов получены результаты:

Предложен алгорим построения компактифицированного горизонтального графа видимости (КГГВ) на снове последовательности дисперсионных оценок слов текста.

Алгоритм апробирован на различных текстах. Оказалось, что для литературных текстов среди узлов, соответствующих КГГВ с наибольшими степенями, присутствуют слова, не только обеспечивающие связность структуры текста, но и определяющие его информационную структуру, отражающие семантику литературных произведений.

1. *Peng C.K. , Buldyrev S.V., Havlin S., Simons M., Stanley H.E., Goldberger A.L.* Mosaic Organization of DNA Nucleotides // Phys. Rev. E 49, 1994. – P. 1685-1689.
2. *Costa M., Goldberger A.L., Peng C.K.* Multiscale entropy analysis of complex ... heartbeat: loss of time irreversibility in aging and disease // *Phys. Rev. Lett.*, 2005. – № 95 (19). – *198102*.
3. *Bezsudnov I.V., Gavrilov S.V., Snarskii A.A.* From time series to complex networks: the Dynamical Visibility Graph // Preprint Arxiv, 2012 (1208.6365). – 13 p.
4. *Albert R., Barabási A.-L.* Statistical machanics of complex networks // Rev. Mod. Phys, 2002. – 74. – P. 47-97.
5. *Dorogovtsev S.N., Mendes J.F.F.* Evolution of Networks. From Biological Nets to the Internet and WWW. – Oxford University Press, Oxford, 2003. – 280pp.
6. *Newman M.E.J.* The structure and function of complex networks // SIAM Rev., 2003. – 45. – P. 167-256.
7. *Zhang J., Small M.* Complex Network from Pseudoperiodic Time Series: Topology versus Dynamics // Phys. Rev. Lett., 2006. – 96. – 238701
8. *Lacasa L., Luque B., Ballesteros F., Luque J., Nuño J.C.* From time series to complex networks: The visibility graph. // Proc. Natl. Acad. Sci. U.S.A., 2008. – 105, 4972-4975.
9. *Luque B., Lacasa L., Ballesteros F., Luque J.* Horizontal visibility graphs: exact results for random time series. // Phys. Rev. E, 2009. – 80. – 04103.
10. *Солганик Г. Я.* Синтаксическая стилистика. Сложное синтаксическое целое. – 2-е изд., испр. и доп. – М.: Высш. шк., 1991. – 182 с.
11. *Broder A.* Identifying and Filtering Near-Duplicate Documents, COM'00 // Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching, 2000. – P. 1-10.
12. *Ferrer-i-Cancho R., Sole R. V.* The small world of human language // Proc. R. Soc. Lond., 2001. – B 268, 2261.
13. *Caldeira S. M. G., Petit Lobao T. C., Andrade R. F. S., Neme A., Miranda J. G. V.* The network of concepts in written texts // Preprint Arxiv, 2005 (physics/0508066). – 10 p.
14. *Ferrer-i-Cancho R.* The variation of Zipf's law in human language. // Phys. Rev. E, 2005. – 70, 056135.
15. *Motter A. E., de Moura A. P. S., Lai Y.-C., Dasgupta P.* Topology of the conceptual network of language // Phys. Rev. E, 2002. – 65, 065102(R).
16. *D.V. Lande, A.A. Snarskii.* Compactified Horizontal Visibility Graph for the Language Network // Preprint Arxiv, 2013 (1302.4619). – 9 p.