Применение графов горизонтальной видимости в информационной аналитике

© Ландэ Д.В.
Институт проблем регистрации информации Национальной академии наук Украины,
Киев, Украина
dwlande@gmail.com
© Снарский А.А.
Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского»,
Киев, Украина
asnarskii@gmail.com

Аннотация

В настоящее время получили широкое распространение методы исследования временных рядов, в основе которого лежит их преобразование в сеть. При таком отображении объединяются две развитые области исследований – нелинейные методы анализа временных рядов и методы теории сложных сетей. Появляется возможность применить развитые методы анализа сложных сетей к анализу временных рядов.

В докладе описывается применение метода графов горизонтальной видимости (Horizontal Viability Graph – HVG) в информационной аналитике. Рассматривается применение концепции Horizontal Viability Graph к двум областям информационной аналитики – построению моделей предметных областей из ключевых слов, отражающих наиболее важные понятия в тематических информационных потоках, а также построению сети связей источников информации. При формировании сети устанавливается связь источника информации с другим, более рейтинговым, если он опубликовал информацию раньше. Построенная таким образом сеть отражает связи источников по заданной тематике, позволяет определять лидеров среди них, делать предположения относительно первоисточников информации.

Данный метод обеспечивает значительное сокращение количества связей между узлами, оставляя наиболее значимые. Предложенный подход, в отличие от традиционных, обеспечивает создание более наглядных сетевых структур.

В докладе рассмотрены реализации концепции Horizontal Viability Graph при анализе тематических информационных потоков, которые расширяют сферу применения данного подхода и позволяют решать задачи наглядной визуализации моделей предметных областей как сети терминов, соответствующих основным понятиям и связям между ними, а также формирования сетей взаимосвязей источников информации.

Ключевые слова: граф горизонтальной видимости, поисковая система, сеть слов, модель предметной области, сеть источников информации.

1 Введение

В настоящее время в естественных науках и технологиях получили широкое распространение методы 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. Появляется возможность применить развитые методы анализа сложных сетей [1, 2] к анализу временных рядов. Предложено несколько методов построения сетей на основе временных рядов [3, 4], в частности, так называемый граф горизонтальной видимости (Horizontal Visibility Graph – HVG) [5, 6].

Алгоритм построения графов видимости по временному ряду проиллюстрирован на рис. 1. При построении графа горизонтальной видимости на горизонтальной оси (ось времени) отмечаются точки, соответствующие индексу моменту времени t_i, от которых в перпендикулярном направлении строятся отрезки высотой, равной значениям ряда в этих точках – $x(t_i)$. Узлами графа горизонтальной видимости являются внешние вершины построенных отрезков. Связь между вершинами в HVG считается существующей, если горизонтальная прямая, проведенная из одной из вершин пересекает отрезок другой вершины, не пересекая ни одного из построенных отрезков, находящихся между ними. Этот геометрический критерий можно записать следующим образом: два узла (элемента ряда), например, t_m и t_n соединены связью, если (см. рис. 2) , $x(t_m)$, $x(t_n) > x(t_p)$ для всех p: $n < p < m$.

Работа посвящена применению концепции Horizontal Viability Graph к двум областям информационной аналитики – построению моделей предметных областей из ключевых слов, отражающих наиболее важные понятия в тематических информационных потоках, а также построению сети связей источников информации.

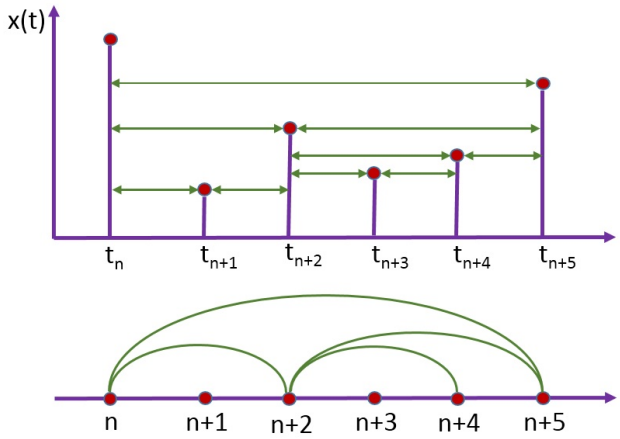


Рис. 1. Пример построения графа горизонтальной видимости

2 Сеть из ключевых слов – модель предметной области

Известно несколько подходов к построению так называемых сетей слов (Language Network) из текстов и способов интерпретации узлов и связей. Например, узлы могут быть соединены между собой, если соответствующие им слова стоят рядом в тексте, принадлежат одному предложению или абзацу, соединены синтаксически или семантически [7, 8]. В [9] был предложен метод построения сети слов путем анализа текстовых корпусов с применением графов видимости.

Поисковая система выдает по тематическому запросу документы, в каждом из которых выделены M ключевых слов, ранжированных в порядке важности, причем ранг определяется весовым значением ключевого слова, например, его относительная частота в документе (слова из стоп-словаря, естественно, не учитываются), TFIDF (TF – частота слова в документе, IDF – величина, обратная количеству документов из массива, в котором встретился данный термин) и его модификации. Рассмотрим последовательность весовых значений ключевых слов из документов по некоторой узкой тематике, определяемых, например, запросом, следующих друг за другом в порядке их появления в информационном потоке (времени публикации). На рис. 2 представлена цепочка документов и характеризующих их ключевых слов и их «видимость справа налево». Предполагается, что первый по времени документ в информационном потоке – D_1, за ним следуют D_2, D_3 и т.д. В пределах одного документа ввиду ранжированности по весу соседние ключевые слова связываются друг с другом (справа налево), и лишь лидирующие по весу ключевые слова могут связываться с ключевыми словами других документов (не обязательно с самими большими весовыми значениями).

Следует отметить, что одно и то же ключевое слово может присутствовать в разных документах и иметь в них различные весовые значения. При этом одно и то же ключевое слово в каждом документе может «видеть» различные ключевые слова, т.е. выходная степень его как узла графа горизонтальной видимости может превышать 1. Связь между некоторыми лидирующими по весу ключевыми словами из различных документов может ассоциироваться с «клубом богатых» [10, 11], феномену, присущему большинству сетей языка [12]. Традиционно связи в сетях языка строятся путем соединения тех узлов (ключевых слов), которые соответствуют одному фрагменту текста (документу, абзацу или предложению), т.е. в рассматриваемом примере степень каждого из узлов не может быть меньше M.

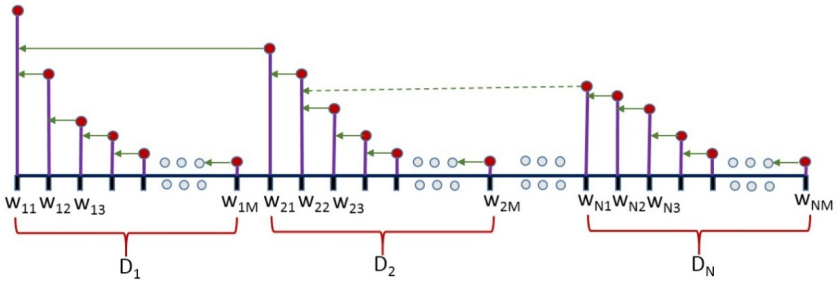
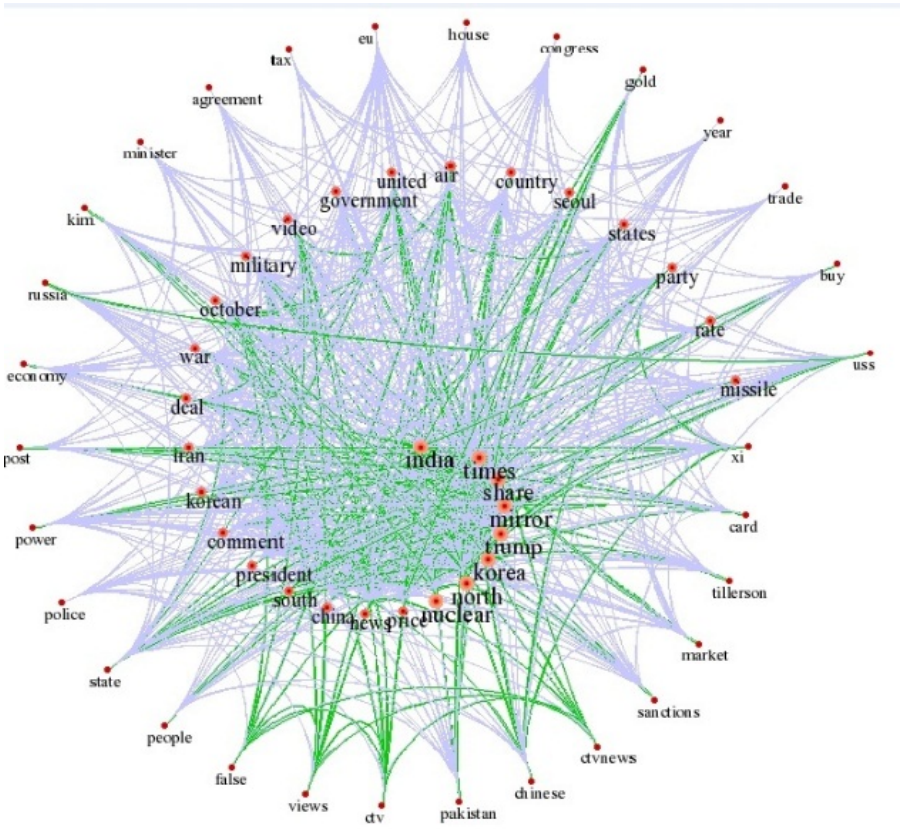
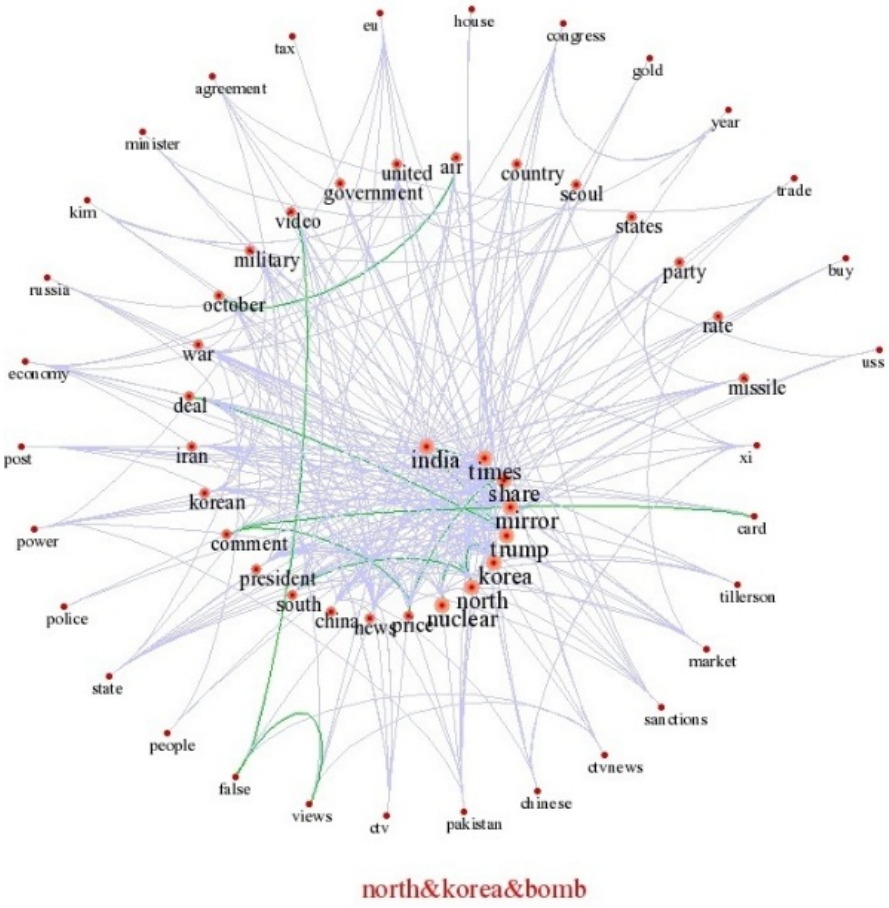


Рис. 2. Горизонтальная видимость ключевых слов w_{ij} из документов D_i

Оценить качество сети, построенной в соответствии с концепцией HVG могут лишь пользователи – аналитики информационной системы, но данный метод обеспечивает значительное сокращение количества связей между узлами (рис. 3), оставляя наиболее значимые (между ближайшими по весу, что важно для ассортативных сетей [13, 14], в которых наряду с «феноменом клуба богатых» существенны связи между узлами, имеющими близкие степени). На рис. 3 представлены две сети связи терминов, первая – традиционная, а вторая построенная с применение предложенного метода. Как можно видеть, второй подход обеспечивает создание более наглядных сетевых структур.



a)



b)

Рис. 3. Сети языка из ключевых слов: a) традиционная; b) Horizontal Visibility Graph

3 Сеть взаимосвязи источников информации

При анализе тематических информационных потоков из различных источников, формируемых системами контент-мониторинга, существует проблема построения сети связей источников информации. При

этом основанием для связи между двумя источникам может служить тот факт, что они часто публикуют совпадающие или близкие по теме документы. Построение такой сети позволяет определять, какие из информационных источников по данной тематике являются основными, наиболее влиятельными, какие подвержены определенным информационным влияниям.

Предлагаемый подход базируется на предположении, что источники информации заранее ранжированы по объемам публикаций исходя из опыта наблюдения за ними в течение продолжительного времени, т.е. им уже приписаны некоторые весовые значения.

Предполагается построение временного ряда, элементы которого будут соответствовать документам из тематического информационного потока в порядке их появления в системе контент-мониторинга сетевых ресурсов (публикации на веб-ресурсах, в социальных сетях). Значения ряда будут соответствовать весовому значению информационного источника, опубликовавшему соответствующий документ. Предполагается, что весовое значение, степень важности источника определены заранее. По данному ряду, как и в предыдущем случае строится граф горизонтальной видимости, причем «направление взгляда» как и в предыдущем примере направлено в прошлое (рис. 4).

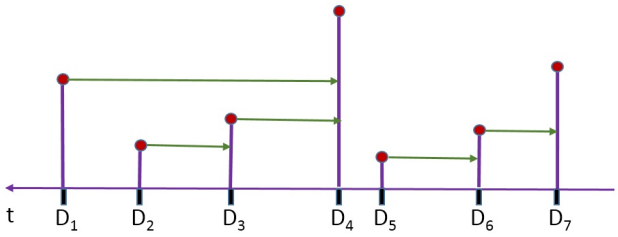


Рис. 4. Горизонтальная видимость документов D_i тематического информационного потока

Таким образом устанавливается связь источника информации с другим, более рейтинговым, если он существует и опубликовал информацию раньше. Построенная таким образом сеть (рис. 5) отражает связи источников по заданной тематике, позволяет определять лидеров среди них, делать предположения относительно первоисточников информации. Например, на рис. 5 показана сеть связей международных источников информации, отражающих тематику ядерных испытаний в Северной Корее в октябре 2017 г. по данным системы контент-мониторинга.

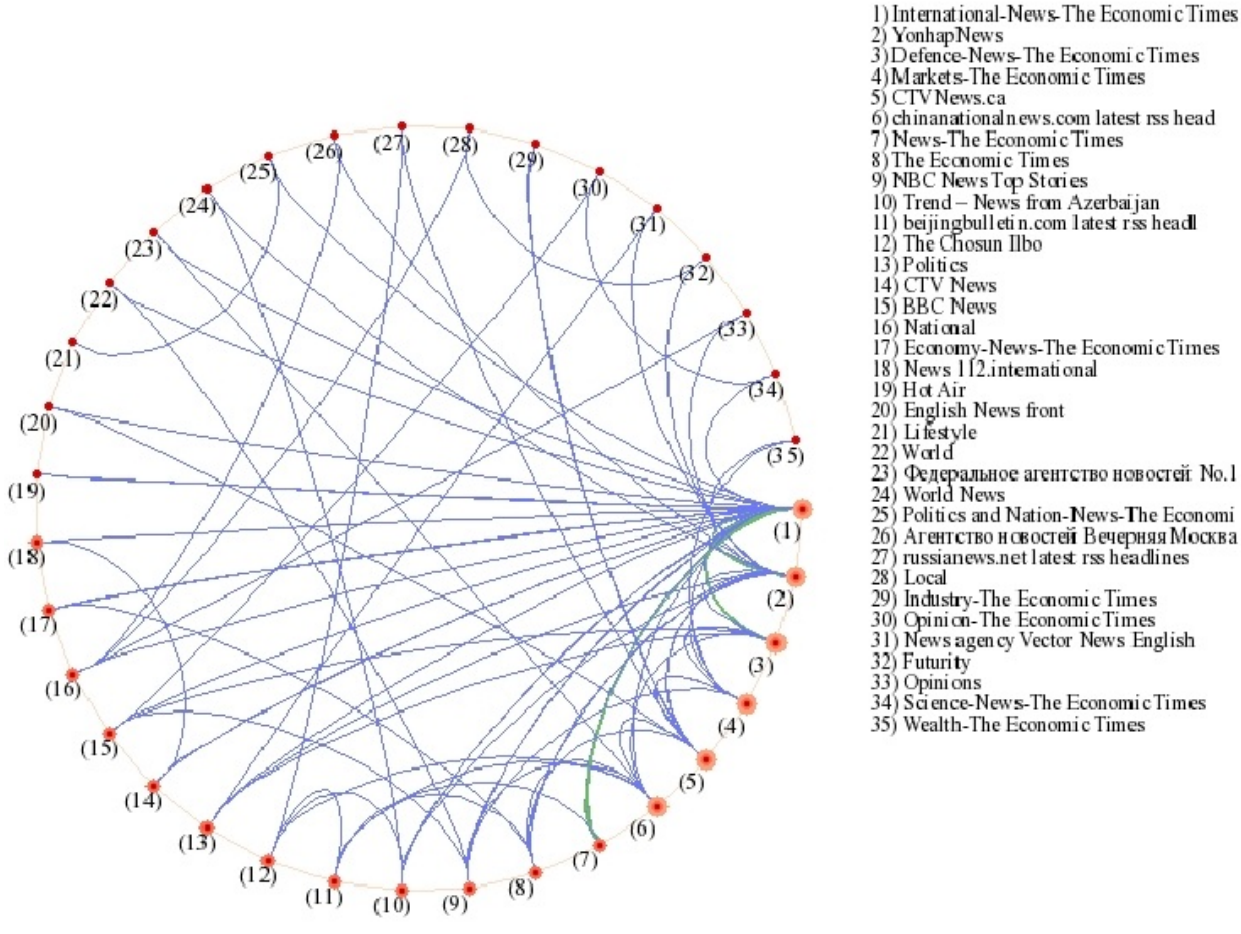


Рис. 5. Пример графа горизонтальной видимости, отражающего связи источников информации по заданной теме

4 Выводы

Предложены реализации концепции Horizontal Viability Graph при анализе тематических информационных потоков, которые, с одной стороны, расширяют сферу применения данного подхода, а, с другой стороны, позволяют решать задачи формирования и наглядной визуализации моделей предметных областей как сети терминов, соответствующих основным понятиям и связям между ними и сетей взаимосвязей источников информации.

Публикация содержит результаты исследований, проведенных при грантовой поддержке Государственного фонда фундаментальных исследований Украины по конкурсному проекту Ф73/23558 «Разработка методов и средств поддержки принятия решений при выявлении информационных операций».

Литература

1. Albert R., Barabási A.-L. Statistical mechanics of complex networks // Rev. Mod. Phys, 2002. – 74. – pp. 47-97.
2. Newman M.E.J. The structure and function of complex networks // SIAM Rev., 2003. – 45. – pp. 167-256.
3. Nunez A. M., Lacasa L., Gomez J. P., Luque B. Visibility algorithms: A short review // New Frontiers in Graph Theory, Y. G. Zhang, Ed. Intech Press, ch. 6. – pp. 119-152 (2012).
4. Bezsudnov I.V., Snarskii A.A. From the time series to the complex networks: The parametric natural visibility graph // Physica A: Statistical Mechanics and its Applications, 2014, 414, 53-60.
5. Luque B., Lacasa L., Ballesteros F., Luque J. Horizontal visibility graphs: Exact results for random time series // Physical Review E,– pp. 046103-1–046103- 11 (2009).
6. Gutin G., Mansour T., Severini S. A characterization of horizontal visibility graphs and combinatoris on words // Physica A, – 390 – pp. 2421-2428 (2011).
7. Ferrer-i-Cancho R., Sole R. V. The small world of human language // Proc. R. Soc. Lond. – B 268, 2261 (2001).
8. 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 physics/0508066 (2005). [6] Ferrer-i-Cancho R., Sole R.V., Kohler R. Patterns in syntactic dependency networks // Phys. Rev. E 69, 051915 (2004).
9. D. V. Lande, A. A. Snarskii, E. V. Yagunova, E. V. Pronoza The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text // 12th Mexican International Conference on Artificial Intelligence, 2013. – pp. 209-215.
10. Colizza, V. and Flammini, A. and Serrano, M. A. and Vespignani, A. Detecting rich-club ordering in complex networks // Nature Physics. 2, 2006. – **2**: 110–115.
11. Julian J. McAuleya, Luciano da Fontoura, Tibério S. Caetan. Program Rich-club phenomenon across complex network hierarchies // Applied physics letters, 2007, -V 91 (8), 084103
12. Heuvel M.P., Sporns O. Rich-Club Organization of the Human Connectome // Journal of Neuroscience, 2011. – 31 (44). – pp. 15775-15786.
13. M.E.J. Newman. Assortative mixing in networks // Phys. Rev. Lett. 89, 208701 (2002).
14. Piraveenan, M., Prokopenko M., and A. Y. Zomaya. Local assortativeness in scale-free networks // EPL (Europhysics Letters) 84.2, 28002 (2008).

Usage of Horizontal Visibility Graphs in Information Analytics

© Dmitry V. Lande
Institute for Information Recording of National Academy of Sciences of Ukraine,
Kyiv, Ukraine
dwlande@gmail.com
© Andrei A. Snarskii
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute",
Kyiv, Ukraine
asnarskii@gmail.com

Abstracts

At present, time series investigating methods based on their transformation into a network are widely used. At this mapping, two developed research domains join - nonlinear methods of time series analysis and methods of the

complex networks theory. Appears an opportunity to apply advanced methods of complex networks analysis for time series.

The paper describes an application of Horizontal Visibility Graph (HVG) method in information analytics. The application of the Horizontal Viability Graph concept for two information analytics domains of is considered: subject domain models building from keywords that reflect the most important concepts in thematic information flows, as well as information sources network building. When forming a network, the information sources link to another, more rated, if it has published information earlier. Constructed in the such way network reflects the connections of sources on a given topic, allows to identify leaders among them and to make assumptions about the primary sources of information.

This method provides a significant reduction of the number of nodes links, leaving the most significant ones. The proposed approach, unlike traditional ones, provides the creation of more demonstrative network structures.

The paper deals with the implementations of the Horizontal Viability Graph concept for analyzing thematic information flows, which expand the scope of this approach usage and allow to solve the tasks of demonstrative visualizing the models of subject domains as a network of terms that correspond to the basic concepts and relationships between them, as well as the formation of information sources relationships networks.

Keywords: horizontal visibility graph, retrieval system, language network, subject domain model, information sources network.