Transformations of Texts into the Complex Network with Applying Visibility Graphs Algorithms

Dmitry Lande[1[0000-0003-3945-1178]], Oleh Dmytrenko[1[0000-0001-8501-5313]] and Andrei Snarskii[2[0000-0002-4468-4542]]

[1] Institute for Information Recording of National Academy of Sciences of Ukraine, Kyiv, Ukraine
[2] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
dwlande@gmail.com, dmytrenko.o@gmail.com, asnarskii@gmail.com

Abstract. In this article, the algorithms of visibility for transforming texts into a complex network is proposed. Keywords and concepts from a set of documents which describe some subject domain are extracted. Numeric values are assigned to each word or phrase using GTF metric, which was proposed in this article instead ordinary TF-IDF metric, that is intended to reflect how important a word is to a document in a collection or corpus. As the result, a time series is constructed. A tool in time series analysis – the visibility graph algorithm is used for constructing a graph of the subject domain. In this article, two actual subject domains ("Information extraction" and "Complex network") are considered for example. The corpora of documents, which are related to actual subject domains, were considered from an open access repository of electronic preprints – arXiv (https://arxiv.org). The proposed algorithm is used for the set of documents, which are related to "Information extraction" and "Complex network". This article shows that applying GTF metric is more expedient compared with TF-IDF metric in the case when the set of documents describe one subject domain. Also, the results of applying the visibility graph algorithm and the compactified horizontal visibility graph algorithm are compared. This article shows, that in some case using the compactified horizontal visibility graph algorithm gives a network of words with more quantity of connections between concepts compared with using the visibility graph algorithm. An open-source visualization and exploration software for all kinds of graphs and networks Gephi and an original package of specially developed Python modules are used for simulation and visualization as an additional tool. The proposed algorithm can be used for visualization some subject domain, and also for information support systems, enabling to reveal key components of a subject domain. Also, the results of this article can be used for building UI of information retrieval systems, enabling to make a process of search a relevant information easier.

Keywords: Set of Documents, Subject Domain, Time Series, Network of Words, TF-IDF, Visibility Graph, Compactified Horizontal Visibility Graph.

1 Introduction

The development of the Internet caused a number of problems, which are related, first of all, with a massive quantity of data in the Web-space, including needless data.

Today on the Internet there is a huge and dynamic information base which is available for research and analysis. It turned out, that many tasks, which arise during working with the network information space, have much in common with mathematical sciences. This fact opens wide opportunities to applying a powerful mathematical tool [1,2]. Taking into account the problems of the huge dimensionality and the dynamic of information resources in global networks, the knowledge based on discrete mathematics (graph theory, networks theory), pattern recognition (classification, clustering), linguistics, digital signal processing, wavelet analysis and fractal analysis are applied.

Due to terabytes of textual data, that are distributed in networks and have been accumulating dynamically, development of new methods and algorithms for analyzing these data is necessary. But also the advantages and disadvantages of algorithms that exist for information retrieval must consider.

A modern development of technologies in some case enable to find relevant information. But the problems of further analytical processing of this information, selection of necessary factual data, detection of development trends in selected subject domain, the relation between concepts, events, and forecasting remain unresolved. More of these problems are actual challenges of a semantic processing of huge dynamical sets of textual data.

2 Analysis of Recent Researches and Publications

A subject of this study is actual and most commonly found in various articles of foreign and domestic scientists. For example, in the works [3,4] the main accent makes on developing new methods and algorithms, which are appointed to analytical processing of huge sets of textual data. In the works [5,6] authors consider a linguistic processing of natural language texts, as one of the central problem of intellectualization of information technologies.

In particular, in the works [7-10] the visibility graphs algorithm is proposed. Also the method of constructing networks based on the visibility graphs algorithm is presented in works [11-15].

3 Review of Some Visibility Algorithms

In this work, a network of connections between terms and concepts, which go into textual data is building. Building networks of words, the nodes of which are elements of the text, enables to reveal key components of the text. At the same time, the task of determining, which of the important structural elements of the text are also informationally important, is actual.

There are several approaches of constructing networks from the texts (so-called language networks) and different ways of interpreting nodes and connections. It leads, accordingly, to various kinds of presenting of such networks. Nodes are connected if corresponding words are either adjacent in the text [16, 17], or are in a single sentence [18], or are syntactically [19, 20] or semantically [21, 22] connected.

3.1 Visibility Graph Algorithm (VG)

In this article, a tool in time series analysis – the visibility graph algorithm [7, 23, 24] is used for converting a time series into a graph. This algorithm maps a time series into a network.

For example, the derived graph of visibility for the time series {0.125, 0.063, 0.042, 0.104, 00.125, 0.063, 0.042, 0.104, 0.125, 0.063, 0.042, 0.104} is presented in **Ошибка! Источник ссылки не найден.**. In the graph, every node corresponds, in the same order, to series data. The visibility rays between the data define the links connecting nodes in the graph.

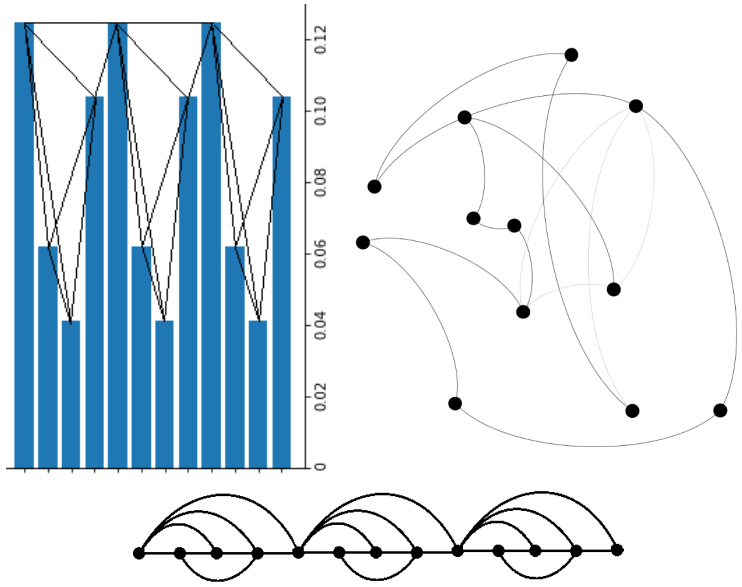

Fig. 1. Example of a time series and the associated graph derived from the visibility algorithm

There is a connection between nodes if they are in “line of sight” with each other, i.e. if they can be connected by a line that does not cross any other histogram bar. More formally, the visibility criteria is described as follows: two arbitrary data values (t_a, y_a) and (t_b, y_b) will have visibility, and consequently will become two connected nodes of the associated graph, if any other data (t_c, y_c) placed between them fulfills:

$$y_c < y_b + (y_a - y_b)\frac{t_b - t_c}{t_b - t_a}.$$

Also in the article [7] is shown that the structure of the time series is conserved in the graph topology: periodic series convert into regular graphs, random series into random graphs, and fractal series into scale-free graphs.

3.2 Compactified Horizontal Visibility Graph Algorithm (CHVG)

In the works [11, 12, 13, 25-27] another algorithm for constructing networks of words – the compactified horizontal visibility graph algorithm (CHVG) is proposed. In general, the process of constructing the language network using the compactified horizontal visibility graph algorithm consists of three stages (Fig. 2). At the first stage, the set of nodes, which correspond to the set of words in order of occurrence in the text, are marked on the horizontal axis. At the second stage, the horizontal visibility graph is built. Two observations made at times t_i and t_j to be connected in a horizontal visibility graph (HVG) if and only if

$$x_k < \min\{x_i, x_j\}$$

for all t_k with $t_i < t_k < t_j$.

At the third stage, the network, that was obtained at the previous stages, is compactified. As the result, the new network of words – the compactified horizontal visibility graph is obtained.

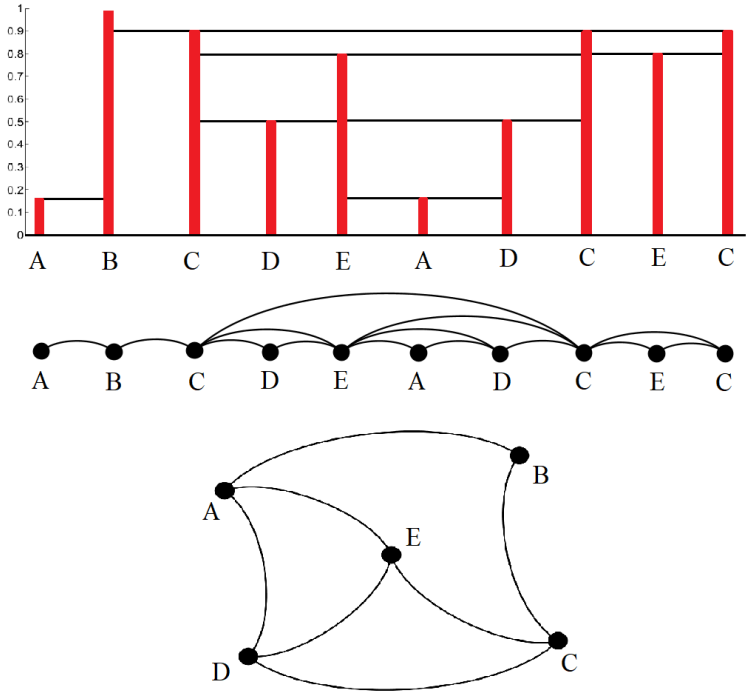

Fig. 2. The stages of the building of the compactified horizontal visibility graph (CHVG)

In this manner, the compactified horizontal visibility graph algorithm enables to construct of network structures based on texts, in which numeric values are assigned in some manner to each word or phrase.

4 Forming of the Time Series

In this article, TF-IDF numeric metric (TF – Term Frequency, IDF — Inverse Document Frequency) is used for forming of the time series. It is an example of a function that assigns a number to a word in the text. TF-IDF is the most frequently applied weighting scheme. Also this a numerical statistic is intended to estimate how important a word is to a document in a collection or corpus [28]. The TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. It is often used as a weighting factor in text mining, information searching, and retrieval. Also, it can be used as one of the criteria to estimate the relevance of a document to a search query [29].

TF (term frequency) is a ratio of the number of the word occurs in a document to the total number of words in the document. In this manner, the weight of a term (word) t_i that occurs in a document is simply proportional to the term frequency. The term was proposed by Karen Spärck Jones [30],

$$TF = \frac{n_i}{\sum_k n_k},$$

where n_i is a number of occurrences of the term (word) i in the document; $\sum_k n_k$ is a total number of words in the document.

IDF (inverse document frequency) is an inverse function of the number of documents in which a term occurs. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term and then taking the logarithm of that quotient. Using IDF reduces the weight of widely used terms (words).

$$IDF = \log \frac{|D|}{|(d_i \supset t_i)|},$$

where $|D|$ is a total number of documents in the corpus; $|(d_i \supset t_i)|$ is the number of documents contain a term t_i ($n_i \neq 0$).

In other words, the TF-IDF metric is a product of two members: TF and IDF.

$$TF - IDF = TF \square IDF$$

A word has high TF-IDF score in a document if it appears in relatively few documents, but appears in this one, and when it appears in a document it tends to appear many times.

After the representation of corpora of documents in a vector view (number of words determines the dimension of the vector), the visibility graph algorithm, which was described above, is used.

5 Presentation of the Basic Material of the Research

In this article, before using the method to constructing networks from the texts, we propose to remove stop word. It enables to removing the words, which are informationally unimportant. We use the stop-dictionary based on various stop-dictionaries, which are available via the links: https://code.google.com/archive/p/stop-words/downloads/; http://www.textfixer.com/tutorials/common-english-words.php.

Also, we propose a global TF metric (GTF), which looks like

$$GTF = \frac{n_i}{\sum_k n_k},$$

where n_i is the number of occurrences of the term (word) i in documents of the corpus; $\sum_k n_k$ is a total number of words in the documents of the corpus.

Words, which are occurred not often within a single document, have a low TF metric. But if they occur in every document of corpora, they at real are informationally important in a global context for the considered subject domain. That is why in this article we use a GTF metric.

5.1 Example 1

In this article, the corpora of 292 documents, which are related with an actual domain – "Information extraction", were considered from an open access repository of electronic preprints – arXiv (https://arxiv.org) for a period of time 2000-2010.

As the result of applying of a proposed method of constructing networks from the texts, the network of keywords, which are important structural elements of the subject domain, was obtained (Fig. 3).

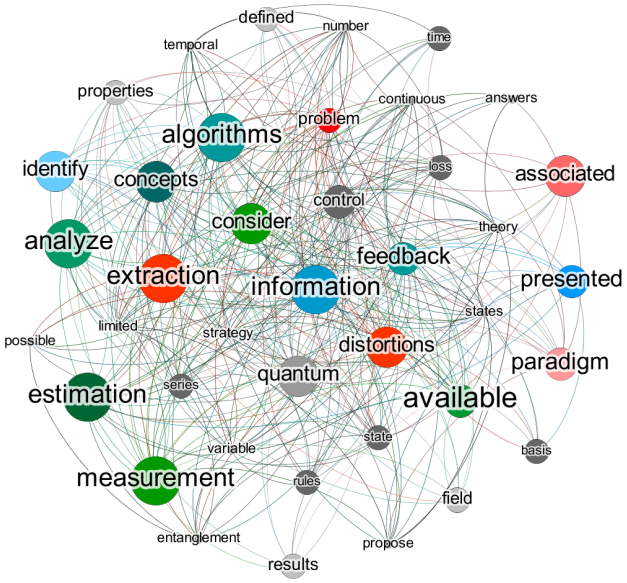

Fig. 3. The network of keywords, obtained through the application of the proposed method

Based on the results which presented in the Table 1 we can notice that quantity of keywords, which are informationally important, is more in case of applying only GTF metric for the set of documents that describe one subject domain. The keywords, such as “information” and “extraction”, which are informationally important for the considered subject domain, are missed in case of using TF-IDF metric (these keywords have a low TF-IDF). After analyzing the results of research (Table 1) we can make the conclusion that applying only GTF metric is more expedient compared with TF-IDF metric in the case when the set of documents describe one subject domain. It can be explained by the fact that words, which are key for the considered subject domain and occur in every document of corpora, have a low IDF (as the result a low TF-IDF). But in fact, these words are informationally important and define the structure of the text.

Table 1. TOP-40 largest-weight nodes of the network of words constructed from corpora of documents, which describe “Information extraction” subject domain

Weight (TF-IDF)	Word	Weight (GTF)	Word
0.23	feedback	0.26	quantum
0.23	quantum	0.26	information
0.22	consider	0.25	feedback
0.218	state	0.239	consider
0.214	control	0.237	control
0.211	measurement	0.205	algorithms
0.21	problem	0.203	problem
0.2	states	0.192	measurement
0.18	continuous	0.186	state
0.179	algorithms	0.185	distortions
0.176	estimation	0.185	concepts
0.176	available	0.178	extraction
0.176	distortions	0.171	loss
0.175	concepts	0.168	limited
0.172	theory	0.166	variable
0.165	identify	0.163	estimation
0.163	limited	0.162	analyze
0.162	paradigm	0.159	states
0.155	defined	0.158	properties
0.152	field	0.158	strategy
0.149	questions	0.157	number
0.1489	considered	0.149	available
0.145	properties	0.143	series
0.145	time	0.141	continuous
0.143	content	0.137	theory
0.14	results	0.136	identify
0.135	entanglement	0.113	entanglement
0.132	presented	0.103	temporal

0.128	number	0.101	propose
0.126	rules	0.092	paradigm
0.124	temporal	0.089	field
0.122	propose	0.085	presented
0.101	label	0.084	rules
0.087	discuss	0.082	defined
0.086	process	0.077	associated
0.083	corresponding	0.072	time
0.076	possible	0.071	possible
0.0745	series	0.07	results
0.053	noise	0.067	basis
0.046	evolution	0.055	answers

5.2 Example 2

For comparison of the results of applying the visibility graph algorithm and the compactified horizontal visibility graph algorithm, the corpora of 2901 documents, which are related with an actual subject domain – "Complex network", were considered from an open access repository of electronic preprints – arXiv (https://arxiv.org) for a period of time 2000-2010. As a result of applying of visibility graph algorithms two different networks of words for the considered subject domain, was obtained (Fig. 4, Fig. 5).

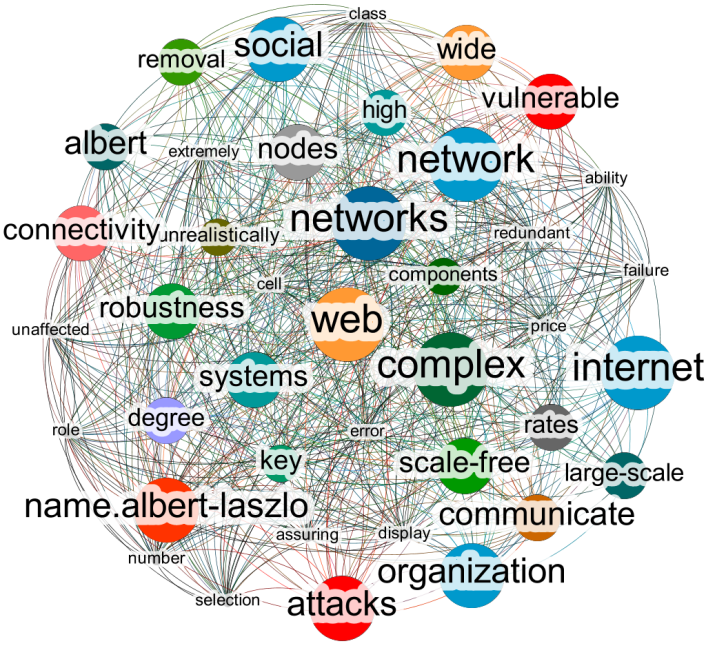

Fig. 4. The network of keywords, obtained through the application of the visibility graph algorithm

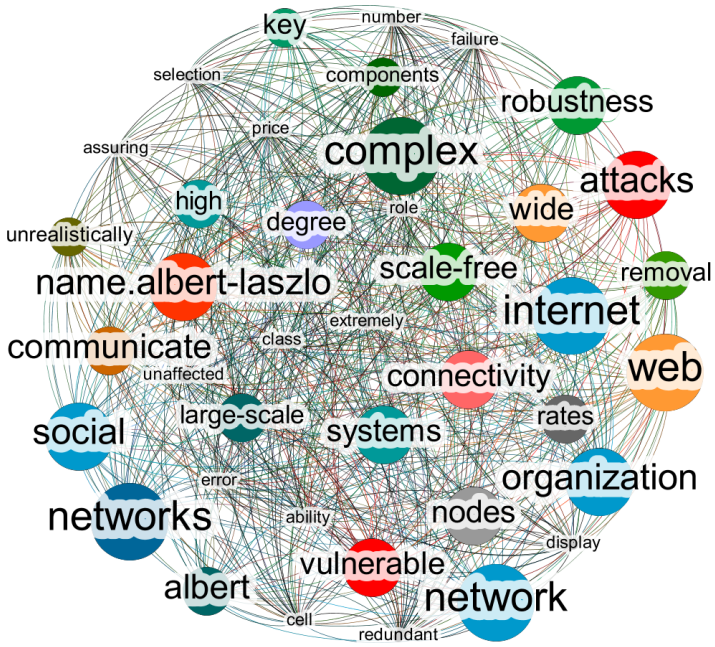

Fig. 5. The network of keywords, obtained through the application of the compactified horizontal visibility graph algorithm

After deriving the associated graphs from the visibility algorithms, all the terms are sorted descending and weight values of CHVG and VG corresponding nodes according to a number of connections with other nodes are calculated. As the weight, for example, the authority (or hub) calculated by HITS algorithm [31] is used. Because the graph is not directed, the choice of a form of the weight does not matter.

Comparing the results (Table 2), it may notice, that in the case of applying the compactified horizontal visibility graph algorithm (Fig. 5) there are many words, which have more links than in the case of applying the visibility graph algorithm (Fig. 4).

Table 2. TOP-40 largest-weight nodes of the network of words constructed from corpora of documents, which describe "Complex network" subject domain

Weight (VG)	Word	Weight (CHVG)	Word
0.17	networks	0.16	selection
0.17	web	0.16	networks
0.17	systems	0.16	removal
0.17	cell	0.16	systems
0.17	key	0.16	nodes
0.17	scale-free	0.16	network
0.17	display	0.16	cell
0.17	error	0.16	ability

0.17	complex	0.16	communicate
0.17	components	0.16	attacks
0.169	nodes	0.16	display
0.169	degree	0.16	rates
0.169	robustness	0.16	robustness
0.168	redundant	0.16	error
0.166	network	0.16	high
0.1645	rates	0.16	unaffected
0.1643	price	0.16	role
0.161	unrealistically	0.16	price
0.16	ability	0.16	connectivity
0.1582	extremely	0.16	extremely
0.158	number	0.158	web
0.1574	high	0.158	degree
0.1573	communicate	0.158	scale-free
0.156	assuring	0.158	wide
0.1547	social	0.158	unrealistically
0.1545	wide	0.157	number
0.1542	internet	0.157	complex
0.1509	unaffected	0.1573	vulnerable
0.1509	class	0.1573	class
0.1505	connectivity	0.1573	internet
0.1504	albert	0.1572	large-scale
0.1502	selection	0.1572	redundant
0.147	attacks	0.1539	social
0.146	removal	0.1539	failure
0.146	vulnerable	0.1534	key
0.145	role	0.1534	name.albert-laszlo
0.138	failure	0.1533	albert
0.135	name.albert-laszlo	0.1533	components
0.1015	large-scale	0.149	assuring
0.1012	organization	0.133	organization

A general quantity of links is 768 in the case of applying the compactified horizontal visibility graph algorithm, unlike in the case of applying the ordinary visibility graph algorithm, when a general quantity of links is 703. It should be noted, that obtained networks are very complex. That is why we plan to continue our research in this sphere.

6 Conclusion

The method of constructing networks from the texts, so-called language networks, was proposed. Keywords and concepts from the set of documents which describe some subject domain were retrieved. Numeric values were assigned to each word or

phrase using GTF metric, which was proposed in this article instead ordinary TF metric. After analyzing the results of the research we made the conclusion that applying only GTF metric is more expedient compared with TF-IDF metric in the case when the set of documents describe one subject domain. As the result, a time series were constructed. A tool in time series analysis – the visibility graph algorithm was used for constructing the graph of the subject domain. After analyzing the results of research the important structural elements of the text were found. It should be noted that these elements of the text also are informationally important and define the structure of the text. There was discovered, that in some case using the compactified horizontal visibility graph algorithm gives a network of words with more quantity of connections between concepts compared with using the visibility graph algorithm. Cause of complexity of obtained networks we plan to continue our research in this sphere.

The proposed method can be used for visualization some subject domain, and also for information support systems, enabling to reveal key components of a subject domain. Also the results of this article can be used for building UI of information retrieval systems, enabling to make a process of search a relevant information easier.

References

1. D.V. Lande, A.A. Snarskii, and I.V. Bezsudnov, Internetika: Navigation in complex networks: models and algorithms, Moscow, Russia: Librokom, Editorial URSS (in Russian) (2009).
2. D.V. Lande, Knowledge Search in INTERNET. Professional work. Dialectics, Moscow (in Russian) (2005).
3. C.C. Aggarwal, and C.X. Zhai, Mining text data. Springer Science & Business Media (2012) 77-128.
4. G. Miner, J. Elder IV, and T. Hill, Practical text mining and statistical analysis for non-structured text data applications. Academic Press (2012).
5. V.Yu. Taranukha, Intelligent processing of texts, Kiev: electronic publication on the website of the faculty (in Ukrainian) (2014).
6. E.I. Bol'shakova, E. S. Klyshinsky, D.V. Lande, A.A. Noskov, O. V. Peskova, and E.V. Yagunova, Automatic processing of texts in a natural language and computational linguistics, Moscow: MIEM Publ (in Russian) (2011).
7. L. Lacasa, B. Luque, F. Ballesteros, J. Luque, and J.C. Nuño, From time series to complex networks: the visibility graph, Proc. Natl. Acad. Sci. USA 105 (2008) 4972–4975.
8. A.M. Nunez, L. Lacasa, J. P. Gomez, and Luque B. Visibility algorithms: A short review, Frontiers in Graph Theory. InTech, (2012) 119 – 152.
9. B. Luque, L. Lacasa, F. Ballesteros, and J. Luque, Horizontal visibility graphs: Exact results for random time series. Physical Review E, 80(4) (2009) 046103.
10. G. Gutin, T. Mansour, and S. Severini, A characterization of horizontal visibility graphs and combinatoris on words, Physica A, – 390 (2011) 2421-2428.
11. D.V. Lande, and A.A. Snarskii, Compactified HVG for the Language Network. In: Proceedings of the International Conference on Intelligent Information Systems: The Conference is dedicated to the 50th anniversary of the Institute of Mathematics and Computer Science, 20-23 Aug. 2013, Chisinau, Moldova: Proceedings IIS, Institute of Mathematics and Computer Science (2013) 108–113.
12. D.V. Lande, A.A. Snarskii, E.V. Yagunova, and E. Pronoza, The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text. In:

Proceedings of the 12th Mexican International Conference on Artificial Intelligence (2013) 209–215
13. D.V. Lande, A.A. Snarskii, and E.V. Yagunova, Application of the CHVG-algorithm for scientific texts. In: Proceedings of the Open Semantic Technologies for Intelligent Systems (OSTIS), February 20 – 22th, Minsk (2014) 199–204
14. D.V. Lande, A.A. Snarskii, and D.Yu. Manko, The Model of Words Cumulative Influence in a Text. In: XVIII International Conference on Data Science and Intelligent Analysis of Information. Springer, Cham (2018) 249-256.
15. D.V. Lande, A.A. Snarskii, E.V. Yagunova, E. Pronoza, and S. Volskaya, Hierarchies of Terms on the Euromaidan Events: Networks and Respondents Perception, 12th International Workshop on Natural Language Processing and Cognitive Science NLPCS 2015 127-139.
16. R. Ferrer-i-Cancho, and R.V. Solé, The Small World of Human Language, Proceedings of the Royal Society of London B: Biological Sciences 268.1482 (2001) 2261-2265.
17. S.N. Dorogovtsev, and J.F.F. Mendes, Language as an Evolving Word Web, Proceedings of the Royal Society of London B: Biological Sciences 268.1485 (2001) 2603-2606.
18. S.M.G. Caldeira, T.C. Petit Lobao, R.F.S. Andrade, A. Neme, and J.G. Miranda, The network of concepts in written texts, Preprint physics/0508066 (2005).
19. R. Ferrer-i-Cancho, R.V. Solé, and R. Kohler, Patterns in syntactic dependency networks, Physical Review E 69.5 (2004) 051915.
20. R. Ferrer-i-Cancho, The variation of Zipf's law in human language, The European Physical Journal B-Condensed Matter and Complex Systems, (2005) 249-257.
21. A.E. Motter, A.P.S. De Moura, Y.C. Lai, and P. Dasgupta, Topology of the conceptual network of language, Physical Review E, 65(6) (2002) 065102.
22. M. Sigman, and G.A. Cecchi, Global Organization of the Wordnet Lexicon, Proceedings of the National Academy of Sciences 99.3 (2002) 1742-1747.
23. I.V. Bezsudnov, and A.A. Snarskii. From the time series to the complex networks: The parametric natural visibility graph, Physica A: Statistical Mechanics and its Applications 414 (2014) 53-60.
24. X. Li, M. Sun, C. Gao, D. Han, and M. Wang, The parametric modified limited penetrable visibility graph for constructing complex networks from time series, Physica A: Statistical Mechanics and its Applications, 492 (2018) 1097-1106.
25. M. Wang, H. Xu, L. Tian, and H. E. Stanley, Degree distributions and motif profiles of limited penetrable horizontal visibility graphs. Physica A: Statistical Mechanics and its Applications (2018).
26. M. Wang, A.L. Vilela, R. Du, L. Zhao, G. Dong, L. Tian, and H. E. Stanley, Exact results of the limited penetrable horizontal visibility graph associated to random time series and its application. Scientific reports, 8(1) (2018) 5130.
27. M. Wang, A.L. Vilela, R. Du, L. Zhao, G. Dong, L. Tian, and H. E. Stanley, Topological properties of the limited penetrable horizontal visibility graph family, Physical Review E, 97(5) (2018) 052117.
28. J.D. Ullman, Data Mining, Mining of massive datasets. Cambridge University Press. (2011) 1–17.
29. J. Beel, B. GIPP, S. Langer, and C. Breitinger, Research-paper recommender systems: a literature survey, International Journal on Digital Libraries. 17(4), (2016) 305-338.
30. K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, MCB University Press 60, (2004) 493-502.
31. J.M. Kleinberg, Authoritative sources in a hyperlink environment. Journal of the ACM JACM. 46 (5) (1999) 604–632.