Creating Directed Weighted Network of Terms Based on Analysis of Text Corpora

Dmytro Lande
The specialized modeling tools department
Institute for Information Recording of National Academy of Sciences of Ukraine
Kyiv, Ukraine
dwlande@gmail.com

Oleh Dmytrenko
Institute for Information Recording of National Academy of Sciences of Ukraine
Kyiv, Ukraine
dmytrenko.o@gmail.com

***Abstract*—One of the most urgent problems of natural language processing – formalization and creation ontological models of subject domains based on the text corpora is considered. In this work, a new approach for determining the weights of links in the network of terms that correspond to certain concepts of the considered subject domain is considered. In particular, applying the proposed approach for determining the weights of links in the network of terms, the terminological ontology of the subject domain that related to an ecological footprint was created as an approbation. Further analysis of the created model made it possible to determine the most influential and significant links between the corresponding nodes in the network of terms that in turn correspond to certain concepts of the considered subject domain. The proposed and considered approaches and methods were programmed and using the software for modeling and visualization of graphs – Gephi the built directed networks of terms were visualized for better visual perception. The weighted directed networks of terms built according to the proposed approach can be used for automatically creating terminological ontologies of subject domains with the participation of experts. Also, the research result can be used to create personal search interfaces for users of information retrieval systems and also can be used in navigation systems in data-bases. It should help users of such systems simplify the process of searching the relevant information.**

Keywords—information space, information flow, text corpus, terminological ontology, subject domain, horizontal visibility graph, undirected networks of terms, directed weighted networks of terms

I. Introduction

The term «information space» usually means the totality of results of semantic human activity, which is presented in the form of information Internet resources where the main results of its communication activity are concentrated. The modern information space is rapidly evolving. As a result, there are huge amounts of data that dynamic information flows contain distributed on the Internet.

But not always from the massive information flows and huge amounts of data it is possible to extract the necessary information that the user needs in response to his request. In particular, it arises because such flows contain a lot of unnecessary data and noise.

As it turned out, and it is confirmed in practice, that there are many problems, that arise when working with the network information space, which has much in common with the mathematical sciences. This fact opens wide opportunities to apply a powerful mathematical tool [1-4].

Because petabytes of textual data are accumulated in information repositories distributed in the network, new approaches and methods of collecting and processing this data to ensure retrieving the information placed on the network are required. Of course, the advantages and disadvantages of existing models and algorithms for information retrieval and analysis should be considered [5–9].

The modern development of technologies allows in some cases to find the necessary information in webspace. But the problems of further analytical processing of this information, the extraction of the necessary factual data, the identification of trends in particular subject domains, the interconnections of objects, events, the recognition of meaningful anomalies, forecasting, and so on, remain. Most of these problems are topical issues of the semantic processing of super-large dynamic text corpora.

Also, the task of content monitoring, as an adaptation of the conception of text mining and classical methods of content analysis to the conditions of formation and development of dynamic information arrays, in particular, information flows on the Internet is a relevant and still unsolved scientific and practical problem. It is because the process of computer processing is usually complicated and requires expensive hardware and software resources and computing systems that most modern servers do not have.

This work considers a new approach for determining the weight of links in the network of terms that correspond to certain concepts of the considered subject domain. It is shown that the proposed approach makes it possible to determine the most influential and significant links between the corresponding nodes in networks of terms that, in turn, correspond to certain concepts of the considered subject domain.

II. Computerized Processing of Text Corpus

There are different techniques for computerized processing and analysis of text as a form of natural language. Given the continuous and rapid development of the information space, there is a need to improve existing approaches and solutions to work with it or to develop new, more adapted to modern trends. In particular, modern software that is designed for processing and analyzing network information flows and arrays also requires advanced. It should be noted that the task of extracting individual terms from the corpus of text documents and automating such extraction is still open and completely unresolved.

One of the main and initial preparatory stages of content analysis is the selection, in particular, of test sources and

978-1-7281-9083-9/20/$31.00 ©2020 IEEE

analysis materials — information retrieval, identification, and formation of the corpus of text documents containing material on a predefined topic or theme.

Also, the stage of processing the formed text corpus is important. The following basic steps of the computerized process of word processing are used in this work: tokenization, lemmatization, stop-words removal, stemming [10, 11], and weighting of terms.

In this work, the stop dictionary was formed from various stop dictionaries that accessible at [12, 13].

The steps described above make it possible to normalize the text of the corpus.

After the stages of the text corpus processing, the weighting and extracting of the key terms are carried out. To form a time series, this study uses the modification of classic statistical weight indicator TF-IDF (TF is Term Frequency, IDF is Inverse Document Frequency) [14, 15] — GTF (Global Term Frequency) [16] as a weight of terms is used to reflect the term to number.

This approach allows informationally-important in global context elements of the text having a high statistical indicator of importance.

III. Determining the Directions of Links

Due to the complexity of natural language, the determination of the syntax and semantic links between nodes that correspond to the terms in the text and the determination of the directions of these links is also an equally complex and open problem of conceptualization.

In this work, a new approach for determining the directions of links between nodes in the undirected network of terms created from words and phrases of a text corpus is presented.

This work considers and applies a modification of a common visibility algorithm that maps a time series into a network – the modification of the Horizontal Visibility Graph (HVG) algorithm — Directed Horizontal Visibility Graph algorithm (DHVG) [17] for creating the directed network of terms as a terminological ontology of some subject domain. The HVG algorithm is an extension of a common visibility algorithm – the Visibility Graph algorithm (VG) [21].

The process of creating an undirected network of terms using the horizontal visibility algorithm consists of two steps [22]. The first step is to mark on the horizontal axis a number of nodes, each of which corresponds to the terms in the order in which they occur in the text; and the weighted values — numerical estimates x_i that is intended to reflect how important a word is to a document in a collection or corpus are marked on the vertical axis. In the second stage, the horizontal visibility graph is created. Two nodes t_i and t_j corresponding to the elements of the time series x_i and x_j, are is connected in a HVG if and only if, when $x_k < \min(x_i; x_j)$ for all t_k ($t_i < t_k < t_j$).

The obtained undirected network of terms is called the horizontal visibility graph (HVG) (see fig. 1). In the fig. 1 the labels «A», «B», …, «E» of nodes are some names of the key terms in the order of their occurrence in the text.

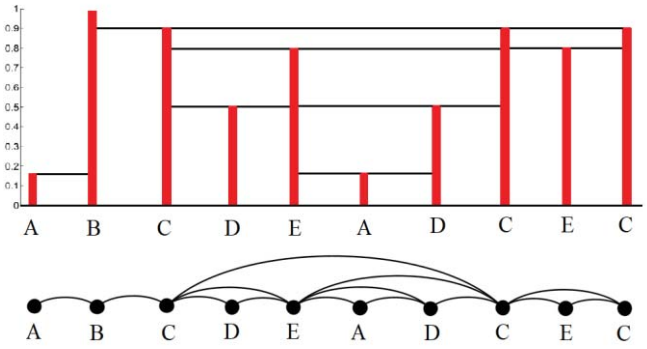

Fig. 1. Stages of creating of the horizontal visibility graph [22]

Using the HVG algorithm makes it possible to create an undirected network of terms from time series that are formed with separated words or phrases of a text corpus and their frequency characteristics.

Next, let's determine the directions of links between nodes in the undirected network of terms according to the following approach. In the undirected network of terms, $G:=(V,T)$ (where V is the set of nodes that correspond to the terms and T is the set of the unordered pairs of nodes from the set V) for $\forall_{i,j}$: $(t_i, t_j) \in T$ link exists in the direction from t_i to t_j if the term defining by the node t_i appearances in some sentence earlier then the term dedined by t_j [17] .

IV. Determining the Weight of Links

Also, the problem of determining the weight of the links in the network of terms is equally complex and open.

In this work, a new approach for determining the weight of links between nodes in the directed network of terms of a text corpus using the algorithm described above is presented.

We can describe the main principle using graph theory terms. Let $D:=(V,E)$ is directed graph that defines the directed network of terms, where V is the set of nodes, E is the set of the ordered pairs of nodes from V. And A is the $N \times N$ adjacency matrix, where $a_{ij} = 1$ if there exists an edge from node i to node j, and $a_{ij} = 0$, otherwise. The nodes of the directed network of terms that correspond to the same terms in the text are merged into a single one. Then to determine the weighted values of the links it needs to concatenate the columns a_{ik} and rows a_{kj} ($1 \le k \le m$) that correspond to the same terms defined by the set $T=\{t_1,..,t_m\}$ (where $1 \le m \le n$). The process described above looks like a weighted compactification of the horizontal visibility graph [22].

A new resulting matrix W will contain the elements w_{ij} which values equal to the number of edges from node i to node j or, in the other words, to the number of occurrences of the term i before the term j in the sentences of the text corpus.

As a result of concatenation, the obtained resulting matrix W defines a directed weighted graph formed of nodes that correspond to the unique terms of the corpus.

V. The Main Results

The proposed approach for determining the direction and weight of the links in the undirected network of terms was tested on the example of the text corpora that thematically related to an ecological footprint.

For the research, a web search engine – Google Scholar [23] was used. At this stage, the annotations of the first 460 articles were downloaded at the query of the «ecological footprint».

According to the stage described above the processing of the text corpora and extracting of the key terms (Table 1) were made.

TABLE I. TOP 34 KEYWORDS FOR THE CORPUS THAT THEMATICALLY RELATED TO ECOLOGICAL FOOTPRINT

№	Word	Weight	№	Word	Weight
1	ecolog	0,0804	18	consumpt	0,0070
2	footprint	0,0716	19	china	0,0069
3	sustain	0,0221	20	product	0,0068
4	develop	0,0143	21	model	0,0066
5	environment	0,0136	22	region	0,0064
6	analysi	0,0126	23	capac	0,0063
7	resourc	0,0109	24	econom	0,0061
8	ef	0,0094	25	account	0,0058
9	assess	0,0093	26	energi	0,0051
10	method	0,0083	27	urban	0,0050
11	human	0,0082	28	nation	0,0049
12	impact	0,0081	29	evalu	0,0047
13	base	0,0081	30	provinc	0,0046
14	indic	0,0074	31	case	0,0046
15	measur	0,0074	32	global	0,0044
16	natur	0,0071	33	tourism	0,0043
17	calcul	0,0071	34	citi	0,0041

Using the Gephi software [24, 25], the directed weighted network of terms was created and visualized (Fig. 2). Fig. 2 depicts the keywords of the considered subject domain.

Also, using the Gephi software tools, the following parameters of the created network were obtained: the number of nodes is 34; the number of links is 778; the average clustering coefficient is 0.717; the average path length is 1.307; the net-work density is 0.693; the number of connected components is 1; the average degree is 22.882.

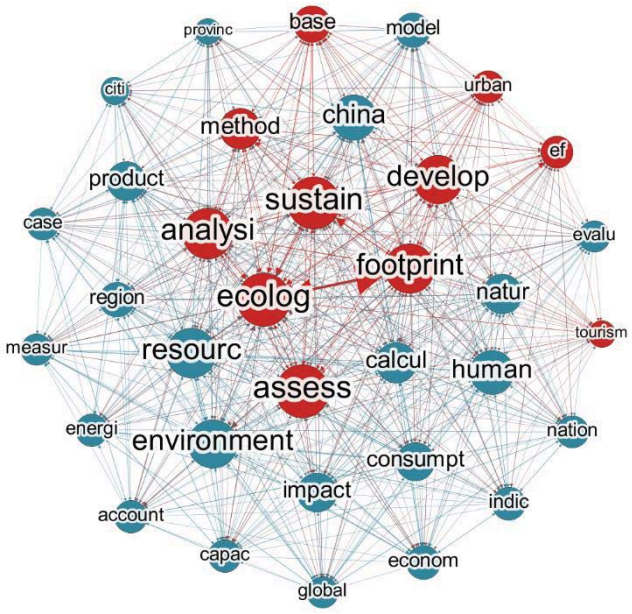

Fig. 2. The directed weighted network of terms created for the subject domain «ecological footprint»

Table 2 shows the list of the most influential and significant links between the corresponding nodes in the network of terms that, in turn, correspond to certain concepts of the considered subject domain.

TABLE II. TOP 29 SIGNIFICANT LINKS FOR THE CORPORA «ECOLOGICAL FOOTPRINT»

№	Source	Target	Weight
1	ecolog	footprint	688
2	footprint	ecolog	494
3	footprint	sustain	176
4	sustain	ecolog	127
5	footprint	analysi	108
6	footprint	environment	94
7	footprint	ef	83
8	develop	ecolog	82
9	analysi	ecolog	81
10	sustain	develop	80
11	environment	ecolog	68
12	base	ecolog	68
13	footprint	resourc	55
14	resourc	ecolog	54
15	footprint	assess	50
16	method	ecolog	47
17	footprint	develop	47
18	footprint	method	47
19	calcul	ecolog	46
20	assess	ecolog	41
21	footprint	account	41
22	footprint	calcul	39
23	footprint	china	36
24	footprint	human	35
25	natur	resourc	33
26	china	ecolog	33
27	ecolog	capac	32
28	ef	ecolog	32
29	human	ecolog	31

After analyzing the obtained results, it was established that the most significant links between the corresponding nodes in the network of terms created for the subject domain «ecological footprint» are: «ecolog → footprint», «footprint → ecolog», «footprint → sustain», «sustain → ecolog», «footprint → analysi» and «footprint → environment».

VI. CONCLUSION

In this work, the main approaches and techniques for computerized processing and analysis of text were considered and applied.

Applying a new approach for determining a direction and weight of the links in the undirected network of terms the ontological model for subject domain related to ecological footprint was created and after detailed analyzing the obtained results the most significant links between the corresponding nodes in the network of terms that, in turn, correspond to

certain concepts of the considered subject domain were established.

The weighted directed networks of terms built according to the proposed approach can be used for automatically creating terminological ontologies of subject domains with the participation of experts. Also, the research result can be used to create personal search interfaces for users of information retrieval systems and also can be used in navigation systems in databases. It should help users of such systems simplify the process of searching the relevant information.

In this work, only statistical methods were used to analyze texts. However, the authors plan to present the results of an analysis using broader natural language processing, such as a part-of-speech tagging, syntactical analysis, and other types of linguistic analysis.

REFERENCES

[1] J. Scott, "Social network analysis. Sociology", vol. 22(1), pp. 109-127, 1988.

[2] D. Knoke, and S. Yang, "Social network analysis" Sage Publications, vol. 154, 2019.

[3] S. P. Borgatti, A. Mehra, D. J. Brass, and Labianca, G., "Network analysis in the social sciences. science", vol. 323(5916), pp. 892-895, 2009.

[4] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding", in Proceedings of the 24th international conference on world wide web, 2015, pp. 1067-1077

[5] U. Brandes, "Network analysis: methodological foundations", Springer Science & Business Media, vol. 3418, 2005.

[6] P. J. Carrington, J. Scott, and S. Wasserman, "Models and methods in social network analysis", Cambridge university press, vol. 28, 2005.

[7] S. K. Bharti, K. S. Babu, A. Pradhan, S. Devi, T. E. Priya, E. Orhorhoro, O. Orhorhoro, V. Atumah, E. Baruah, and P. Konwar, "Automatic keyword extraction for text summarization in multi-document e-newspapers articles", European Journal of Advances in Engineering and Technology, vol. 4(6), pp. 410-427, 2017.

[8] A.A.L. Below, Information retrieval data structures and algorithms, 1992.

[9] C. D. Manning, P. Raghavan, and H. Schütze, "An Introduction to Information Retrieval", Cambridge University Press, pp. 22–36, 2009.

[10] J. B. Lovins, "Development of a stemming algorithm", Mech. Translat. & Comp. Linguistics, vol. 11(1-2), pp. 22-31, 1968.

[11] B. Jongejan, and H. Dalianis, "Automatic training of lemmatization rules that handle morphological changes in pre-, in-and suffixes alike", in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Association for Computational Linguistics, Singapore, 2009, pp. 145-153.

[12] Google Code Archive: Stop-words, https://code.google.com/archive/p/stop-words/downloads/, last accessed 2020/03/11.

[13] Text Fixer: Common English Words List, http://www.textfixer.com/tutorials/common-english-words.php, last accessed 2020/03/11.

[14] G. Salton, and C. Buckley, "Term-weighting approaches in automatic text retrieval", Information processing & management, vol. 24(5), pp. 513-523, 1988.
doi:10.1016/0306-4573(88)90021-0

[15] J. Ramos, "Using tf-idf to determine word relevance in document queries", in Proceedings of the first instructional conference on machine learning, vol. 242, 2003, pp. 133-142.

[16] D.V. Lande, O.O. Dmytrenko, and A.A. Snarskii, "Transformation texts into complex network with applying visability graphs algorithms", in: CEUR Workshop Proceedings (ceur-ws.org). Vol-2318 urn:nbn:de:0074-2318-4. Selected Papers of the XVIII International Scientific and Practical Conference on Information Technologies and Security (ITS 2018). vol. 2318, 2018, pp. 95-106.

[17] D.V. Lande, O.O. Dmytrenko, and O.H. Radziievska, "Determining the Directions of Links in Undirected Networks of Terms", in: CEUR Workshop Proceedings (ceur-ws.org). Vol-2577 urn:nbn:de:0074-2318-4. Selected Papers of the XIX International Scientific and Practical Conference "Information Technologies and Security" (ITS 2019). vol. 2577, 2019, pp. 132-145. ISSN 1613-0073

[18] B. Luque, L. Lacasa, F. Ballesteros, and J. Luque, "Horizontal visibility graphs: Exact results for random time series", Physical Review E, vol. 80(4), 2009.
doi: 10.1103/PhysRevE.80.046103

[19] G. Gutin, T. Mansour, and S. Severini, "A characterization of horizontal visibility graphs and combinatorics on words", Physica A: Statistical Mechanics and its Applications, vol. 390(12), pp. 2421-2428, 2011.
doi: 10.1016/j.physa.2011.02.031

[20] 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, vol. 414, pp. 53-60, 2014.
doi: 10.1016/j.physa.2014.07.002

[21] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, and J. C. Nuno, "From time series to complex networks: The visibility graph", Proceedings of the National Academy of Sciences, vol. 105(13), 2008, pp. 4972-4975.
doi: 10.1073/pnas.0709247105

[22] D.V. Lande, A.A. Snarskii, E.V. Yagunova, and E.V. Pronoza, "The use of horizontal visibility graphs to identify the words that define the informational structure of a text", in: 2013 12th Mexican International Conference on Artificial Intelligence, 2013, pp. 209-215.

[23] Google Scholar, https://scholar.google.com, last accessed 2020/03/11.

[24] K. Cherven, "Network graph analysis and visualization with Gephi", Packt Publishing Ltd, 2013.

[25] Gephi, https://gephi.org, last accessed 2020/03/11.