Subject Domain Models of Jurisprudence According to Google Scholar Scientometrics Data

Dmytro Lande[1,2,3][0000-0003-3945-1178], Oleh Dmytrenko[1][0000-0001-8501-5313] and Oksana Radziievska[3][0000-0003-3813-3987]

[1] Institute for Information Recording of the National Academy of Sciences of Ukraine, Kyiv, Ukraine
[2] National Technical University "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
[3] Scientific Research Institute for Informatics and Law of the National Academy of Legal Sciences of Ukraine, Kyiv, Ukraine

dwlande@gmail.com dmytrenko.o@gmail.com
radeoksa@gmail.com

Abstract. The paper considers the approach to the network structuring of concepts of the selected subject domain based on the data contained in the Google Scholar online scientometrics documentary resource. We present the methods used in the creation of the subject domain models as networks of the terms of a specific topic, which correspond to the basic concepts within the specified theme. In particular, in this work, we use the networks of natural hierarchies of terms – the algorithm for creating the directed network of words and phrases for thematic text corpora as a terminological model. Based on the freely accessible search engine which indexes the full text of scientific publications – Google Scholar for the thematic queries it was pre-prepared the text corpora. Within the scope of this work, the so-called networks of natural hierarchies of terms are considered for the corpus of scientific articles related to the subject domains of "Criminal Law" and "Copyright Law". The considered in this work processes were automated by using the NLTK library of the Python programing language. The obtained networks of natural hierarchies of terms for "Criminal Law" and "Copyright Law" were visualized and analyzed. The considered techniques of creation of such networks and the implementation of the algorithm for creating the directed networks of terms will contribute to the formation and improvement of the conceptual and terminological apparatus in the legal sphere and the harmonization of national and international legislation.

Keywords: Subject Domain Model, Legal Information, Terminological Model, Text Corpus, Network of Terms, Horizontal Visibility Graph, Network of Natural Hierarchies of Terms, Undirected Network, Directed Network, Criminal Law, Copyright Law.

1 Introduction

It is known that modern information space is characterized by the rapid development of dynamic information streams distributed in the webspace – the Internet. But it is not always made possible to get the necessary information that the user needs in response to their request from the massive information flows and huge amounts of data that they contain. In particular, this arises because such flows contain a lot of unnecessary data and noise. That is why, there is a pressing need for research and development of new methods and approaches of retrieving, analyzing and extracting information, its convenient visual presentation.

The different methods, approaches of computerized text processing and generation terminological ontologies are existed [1, 2, 3]. Despite this, new less resource-intensive solutions for processing a huge flow of data are needed.

The structuration of data distributed in modern networks, the formation of networks of subject domains based on automatically extracted terms will help to simplify the task and will contribute to the development and improvement of the conceptual and terminological framework, particularly in the legal sphere and the harmonization of national and international law.

A very significant step in the comprehensive research of the subject domain is a detailed formalized representation of knowledge suitable for automated processing – the creation of the subject domain models, ontologies, including legal ones. A network of terms, in which nodes correspond to the separate key words and phrases and the edges to the connections between them, can be reviewed and applied as a terminological ontology of some considered subject domain [4].

A particular step of this conceptualization is the determination of key objects (in this case, the creation of vocabulary nomenclatures, thesauri and dictionaries of the terms, specified based on thematic sets of text documents). Efficient selection of specific terms from a text set is an urgent and unsolved issue [5, 6]. Also, it is still an unresolved task is to establish the connections between terms.

Also, the question arises about the further visual presentation of subject domains. One of the domain models can be considered a network of words, the nodes of which correspond to a separate concept, and the edges to the connections between them [7, 8].

2 Methods

The initial stage of the creation of a network of terms associated with a particular subject domain is the formation of a corpus of text documents. In this work, it was used a freely accessible web search engine which indexes the full text of scientific publications – Google Scholar (https://scholar.google.com). At this stage, the annotations of the first 385 articles were downloaded at the query of the “Criminal Law” and the annotations of the first 490 articles at the query of the “Copyright Law”.

2.1 Processing Text Documents and Key Terms Extraction

For preliminary lexical analysis, the basic steps of processing text documents as tokenization [9] (splitting the text into elementary units – tokens), lemmatization [10], stemming process [11], removing stop-words and terms weighting are made.

The NLTK library of Python programing language that, in particular, realize a streamer "PorterStemmer" [12, 13, 14] was used as additional tools for the program realization and automatization of processing text documents.

Also, it should be noted that for the considered subject domain it was used the additional stop-word dictionary formed by experts.

2.2 Terms Weighting

The next step is the terms weighting. As a weight value of terms, the classical numerical statistic TF-IDF (Term Frequency-Inverse Document Frequency) [15] is used for forming a time series that will be transformed into an undirected graph – a network of terms. Although this is not the only approach possible to solve the problem of identifying key terms [16]. The statistical weight indicator is applied to assess the weight of terms in the context of a document that is a part of a collection of documents or a corpus [17]. The weight (importance) of a term is directly proportional to the number of the occurrences of the term in the document and inversely proportional to the number of documents of the collection in which the term occurs. The TF-IDF indicator is used in the tasks of text analysis and information retrieval [18]. The terms of the high frequency within the document and a low document frequency in the whole collection of documents will have a higher TF-IDF weight [19].

Because this study processes documents describing one subject domain, then to prevent the loss of informationally-important elements of text, reference words and phrases (bigrams and trigrams), only the TF indicator was used as a statistical indicator of the importance of a term. It is also used because terms that occur in most documents, in general, have a low IDF indicator (so, the TF-IDF numeric value will be low), while in fact, these terms are key and basis in the context of the corpus that formed for some subject domain. That is why, to avoid a situation that arises when working with a text corpus of a predetermined topic, when an informationally-important term occurs almost every document of the collection and has a low TF weight indicator, the Global TF [20] was used:

$$GTF = \frac{n_i}{\sum_k n_k} \tag{1}$$

where n_i is the total number of occurrences of term i in the corpus; $\sum_k n_k$ is the total number of terms in the corpus of documents.

The Global TF makes it possible to have a high statistical weight for informationally-important terms in a global context.

2.3 Algorithm for Creating Compactified Visibility Graph

One of the algorithms for transforming the time series into a graph is an algorithm for creating networks of terms was proposed – the algorithm for building a Compactified

Horizontal Visibility Graph (CHVG), that was proposed in the work [16]. This algorithm is used for building a terminological model for unique key words and phrases (bigrams and trigrams) of text documents. In general, a network of terms, which using the compactified horizontal visibility algorithm, is created in three stages. In the first stage, a number of nodes in the order in which their corresponding terms appear in the text are marked on the horizontal axis. Next, the weight values (the numerical estimates) of terms are marked on the vertical axis. In the second stage, a graph of horizontal visibility is created [21]. More formally, the creation of the horizontal visibility graph can be represented as follows. Time series elements x_i and x_j, are in horizontal visibility, when $x_k < \min(x_i; x_j)$ for all terms t_k ($t_i < t_k < t_j$) that corresponds them.

Next stage is the compactification the undirected graph that obtained on in the previous steps. This stage implies the merging into a one of all nodes that correspond to the same terms. As a result, the compactified horizontal visibility graph (CHVG) (fig. 1) is obtained.

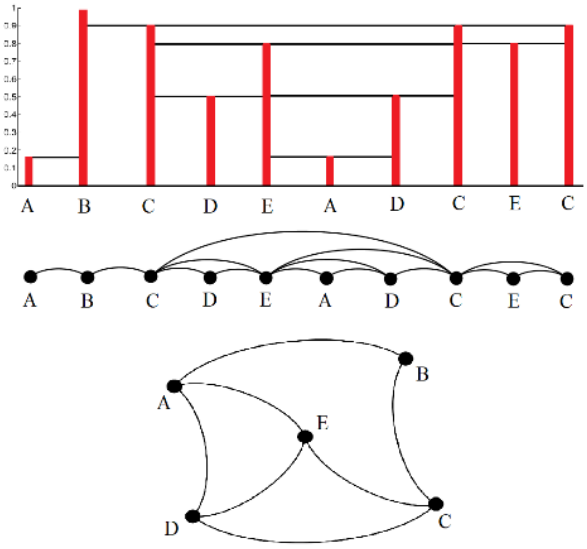


Fig. 1. Three stages of CHVG building [16].

When unique key words, bigrams and trigrams of text corpus have some weight indicator and form time series, then the CHVG algorithm makes it possible to transform it into an undirected network of terms.

2.4 Formation of Network of Natural Hierarchies of Terms

An algorithm for creating a directed network with words and phrases – an algorithm for creating networks of natural hierarchies of terms [16] for a corpus of text documents is one of the possible approaches for creating ontologies of some subject domain. This algorithm is based on the use of information-important elements of the text, reference

words and phrases (bigrams and trigrams), the method for identifying them is presented in this work [16]. The algorithm for creating networks of natural hierarchies of terms includes the creation of a compactified horizontal visibility graph for terms – individual words, bigrams and trigrams, and the establishment of directional relationships between terms.

As indicated in [22], the algorithm for creating networks of natural hierarchies of terms can be represented as successive stages that include the preliminary processing of the text documents set; the selection of keywords and phrases which are information-important within the subject domain; the creation of a compacted horizontal visibility graph (CHVG); recalculation of sorting the weighting values of the selected terms by the specified weight criterion and the selection of the most significant ones. The final stage is the direct formation of a network of natural hierarchies of terms (connecting nodes with "occurrence" links) and its visualization.

For sequences of terms (words, bigrams and trigrams) and their weight values, determined using the statistical term importance indicator GTF, the compactified horizontal visibility graphs (CHVG) are created. The next step is to recalculate the weight values corresponding to the terms in the CHVG. This procedure allows us to consider also those terms that are of great importance for the general subject of the text corpus [20]. In this work, weights are recalculated using the HITS algorithm [23, 24, 25], which determines the authority or hub for each CHVG node. The choice of the weight value form (authority or hub) does not matter since the graph is undirected. After that, all terms are arranged in descending order of the calculated weight values of the corresponding nodes in CHVG.

Further, an expert method determines the required size (number N) of the network of natural hierarchies of terms, that will be created, after which N simple words, bigrams and trigrams (total N+N+N elements) which have the highest weights of the corresponding nodes in CHVG are selected.

At the next step, it is to create the network of the natural hierarchies of terms, where the nodes correspond to the selected terms, and the links between them correspond to the occurrence of one term into the others (Fig. 2).

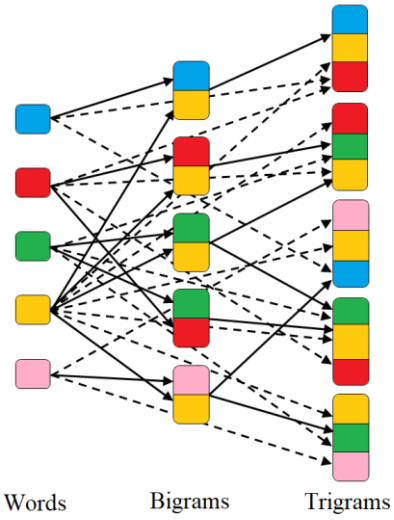


Fig. 2. The three-tier model of the network of natural hierarchies of terms.

The final step is a visualization of the created network of natural hierarchies of terms with tools for graph visualization. The input to such means is the adjacency matrix (in the form of the CSV file) built at the stage of the development of a network of natural hierarchies of terms. The network of natural hierarchies of terms, which created fully automatically, can be applied as a foundation for an automated generation of terminological ontologies with the participation of experts.

3 Visualization and Analysis of Results

In this work, we used the corpus of pre-selected text documents, thematically related to the relevant subject domain – "Criminal Law". Having imported a streamer, which is implemented in Python (NLTK library – Natural Language Toolkit), the process of stemming of the text corpus of 385 documents, obtained at the query of "Criminal Law", was performed; and as a result, the words that have a common root were combined.

Table 1 lists the weightiest terms (words, bigrams and trigrams) for the studied subject domain in accordance with the HITS network rank criterion [23, 24, 25].

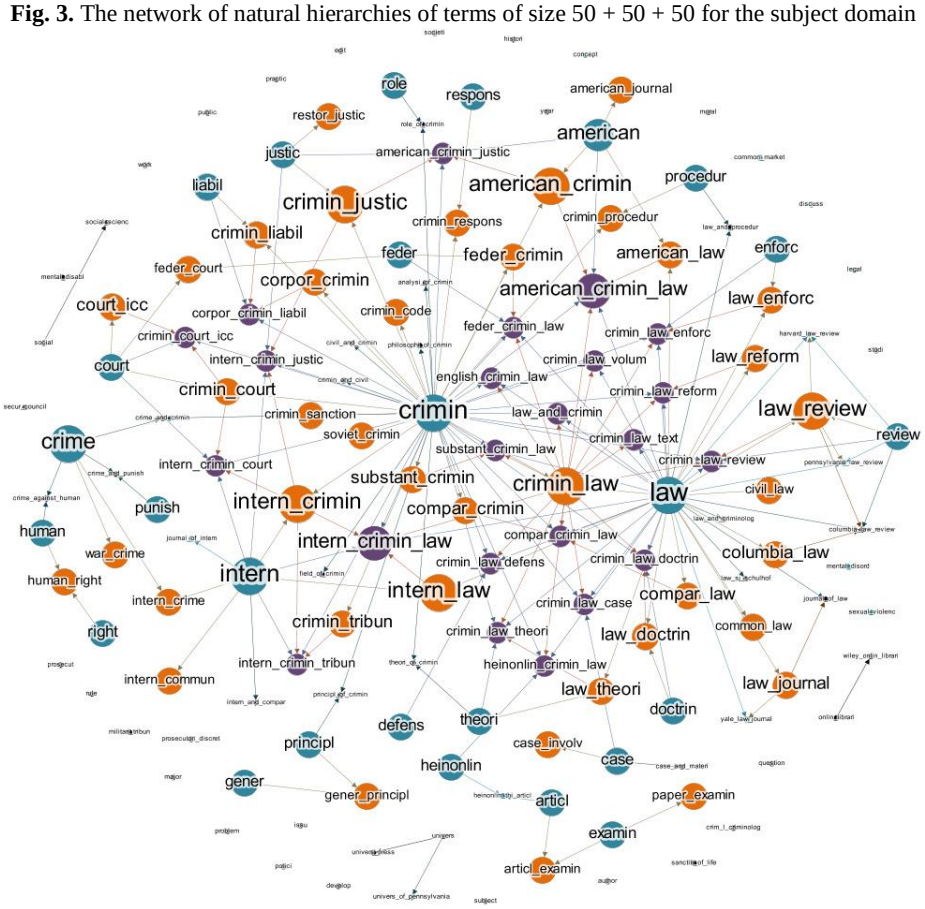
Table 1. Lists of the weightiest terms (words, bigrams and trigrams) for "Criminal Law".

№	words	bigrams	trigrams
1	studi	restor_justic	civil_and_crimin
2	moral	onlin_librari	heinonlin_thi_articl
3	work	law_reform	crimin_law_doctrin
4	law	crimin_sanction	law_and_crimin
5	principl	columbia_law	theori_of_crimin

6	respons	crimin_respons	crimin_law_text
7	subject	civil_law	univers_of_pennsylvania
8	liabil	univers_press	american_crimin_justic
9	human	corpor_crimin	crime_and_crimin
10	right	gener_principl	case_and_materi
11	intern	articl_examin	analysi_of_crimin
12	role	case_involv	intern_and_compar
13	concept	mental_disord	principl_of_crimin
14	gener	compar_crimin	substant_crimin_law
15	univers	court_icc	american_crimin_law
16	examin	crimin_justic	crimin_court_icc
17	practic	feder_crimin	crime_against_human
18	polici	substant_crimin	field_of_crimin
19	year	intern_crimin	sanctiti_of_life
20	theori	crimin_code	philosophi_of_crimin
21	crime	war_crime	columbia_law_review
22	review	common_market	crimin_law_theori
23	case	compar_law	law_sj_schulhof
24	develop	paper_examin	law_and_criminolog
25	legal	sexual_violenc	journal_of_intern
26	heinonlin	law_review	crime_and_punish
27	articl	intern_law	english_crimin_law
28	public	social_scienc	role_of_crimin
29	social	intern_crime	law_and_procedur
30	histori	law_theori	crimin_law_case
31	court	american_law	intern_crimin_justic
32	author	soviet_crimin	wiley_onlin_librari
33	procedur	crimin_liabil	crimin_law_defens
34	american	crimin_procedur	intern_crimin_court
35	major	crimin_tribun	crimin_law_volum
36	discuss	crimin_law	crimin_law_enforc
37	question	feder_court	harvard_law_review
38	prosecut	intern_commun	pennsylvania_law_review
39	crimin	law_journal	crimin_and_civil
40	issu	secur_council	crim_l_criminolog
41	edit	law_enforc	compar_crimin_law
42	societi	law_doctrin	intern_crimin_law
43	defens	american_crimin	crimin_law_review
44	enforc	american_journal	journal_of_law
45	punish	human_right	intern_crimin_tribun
46	doctrin	militari_tribun	yale_law_journal
47	feder	mental_disabl	feder_crimin_law
48	justic	crimin_court	crimin_law_reform
49	problem	common_law	corpor_crimin_liabil
50	rule	prosecutori_discret	heinonlin_crimin_law

Gephi software (https://gephi.org) [26] was used to visualize the network of natural hierarchies of terms in size of 50 + 50 + 50 (Fig. 3).

Fig. 3. The network of natural hierarchies of terms of size 50 + 50 + 50 for the subject domain



"Criminal Law"

Also, using the Gephi software tools, the following parameters of the created network were obtained: the number of nodes is 150; the number of links is 205; the network density is 0.009; the number of connected components is 35; the average path length is 1; the average clustering coefficient is 0.121.

According to the topological singularity, the network has a small average clustering coefficient. It is explained by the presence in the network of a large number of concepts whose neighbors have little connection with each other – this is a sign of the so-called quasi-hierarchical (close to hierarchical, in which the number of connections is comparable to the number of nodes) networks. At the same time, a small average path length indicates that this network is also a “Small World” (Small World) [27].

The corpus of textual documents, thematically related to the relevant subject domain – "Copyright Law" was also processed. The process of the stemming of the corpus of 490 documents obtained at the query of the "Copyright Law" was performed.

Table 2 lists the most weighing terms (words, bigrams and trigrams) for the studied domain in accordance with the HITS network rank criterion.

Table 2. Lists of the weightiest terms (words, bigrams and trigrams) for "Copyright Law".

№	words	bigrams	trigrams
1	digit	copyright_legisl	type_of_tumour
2	intellectu	digit_technolog	wiley_onlin_librari
3	unit	properti_right	author_and_publish
4	music	copyright_protect	heinonlin_thi_articl
5	work	cardozo_art	fair_use_doctrin
6	law	law_reform	german_copyright_law
7	origin	violat_fall	law_in_canada
8	public	canadian_copyright	soc_y_usa
9	fair	three-step_test	access_to_copyright
10	protect	unauthor_copi	case_and_materi
11	properti	legal_studi	law_of_copyright
12	patent	intel_prop	intellectu_properti_right
13	right	intellectu_properti	purpos_of_copyright
14	intern	copyright_handbook	law_and_econom
15	creativ	univers_press	professor_of_law
16	analysi	berkeley_tech	literari_and_artist
17	current	digit_media	digit_millennium_copyright
18	gener	paid_violat	law_jc_ginsburg
19	inform	copyright_law	patent_and_copyright
20	nation	copyright_work	canadian_copyright_law
21	internet	septemb_9	law_is_base
22	theori	copyright_owner	law_a_commentari
23	question	subject_matter	aspect_of_copyright
24	copi	public_domain	public_or_part
25	artist	intern_copyright	librarian_and_educ
26	case	part_thereofispermit	law_of_septemb
27	develop	current_copyright	notion_of_origin
28	author	special_case	european_copyright_law
29	media	copyright_limit	transit_ma_schlosshauer
30	econom	law_review	republ_of_china
31	legal	digit_millennium	analysi_of_copyright
32	industri	american_copyright	intellectu_properti_law
33	creat	exclus_right	paid_violat_fall
34	articl	german_copyright	nation_inform_infrastructur
35	social	copyright_infring	protect_by_copyright
36	court	copyright_case	law_and_practic
37	publish	european_copyright	heinonlin_copyright_law
38	american	unfair_competit	approach_to_copyright
39	number	wto_panel	law_c_geiger

40	art	digit_copyright	version_and_permiss
41	materi	open_sourc	american_copyright_law
42	univers	law_journal	digit_copyright_law
43	copyright	properti_law	current_copyright_law
44	technolog	moral_right	springer-verlag_berlin_heidelberg
45	societi	copyright_fee	berlin_heidelberg_gmbh
46	softwar	copyright_soc	law_a_propos
47	cultur	copyright_issu	law_p_samuelson
48	heinonlin	price_discrimin	intern_copyright_law
49	limit	current_version	dichotomi_in_copyright
50	futur	law_school	guid_to_copyright

A network of natural hierarchies of terms in size of 50 + 50 + 50 for the subject domain "Copyright Law" was built and visualized (Fig. 4).

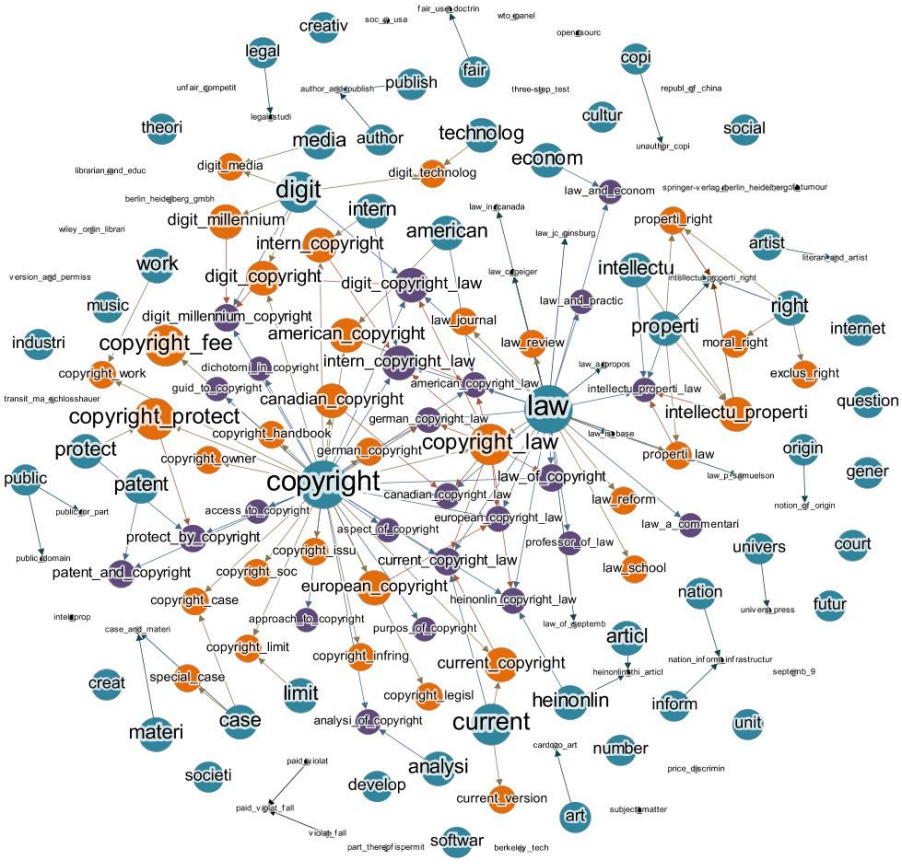


Fig. 4. The network of natural hierarchies of terms of size 50 + 50 + 50 for the subject domain "Copyright Law"

Also, using the Gephi software tools, the following parameters of the created network were obtained: the number of nodes is 150; the number of links is 144; network density is 0.006; the number of connected components is 48; the average path length is 1; the average clustering coefficient is 0.068.

4 Conclusion

The article describes the method for creating a network according to words and phrases - an algorithm for forming networks of natural hierarchies of terms by a set of thematically related text documents. The considered methodology was applied to create the subject domain models "Criminal Law" and "Copyright Law". Based on the freely accessible search engine which indexes the full text of scientific publications – Google Scholar, for the “Criminal Law” and “Copyright Law” queries it was pre-prepared the text corpora of 385 and 490 documents respectively. We obtained the networks of natural hierarchies of terms for a set of text documents thematically related to the "Criminal Law" and "Copyright Law".

The Gephi graph visualization and modeling package and our own set of specially developed modules in the Python programming language were used as auxiliary tools.

So, the fully automated process of creating networks of the natural hierarchies of terms can be used as the foundation for an automated generation of terminological ontologies with the participation of experts. Described in this paper the method for creating a directed network with the words and phrases will contribute to the formation and improvement of the conceptual and terminological apparatus in the legal sphere and the harmonization of national and international law.

References

1. Snarsky, A., & Lande, D.: Modeling of complex networks: a training manual. p. 212. K.: "Engineering" (2015). ISBN 978-966-2344-44-8
2. Below, A.A.L.: Information retrieval data structures and algorithms. (1992).
3. Navigli, R., Velardi, P., & Gangemi, A.: Ontology learning and its application to automated terminology translation. IEEE Intelligent systems, 18(1), 22-31. (2003).
4. Lande, D., & Snarsky, A.: Approach to Creation of Terminological Ontologies. Design ontology 2(12), pp. 83-91. (2014). (in Russian)
5. Aharon, M., Elad, M., & Bruckstein, A.: K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on signal processing, 54(11), 4311-4322. (2006).
6. Gilchrist, A.: Thesauri, taxonomies and ontologies–an etymological note. Journal of documentation. (2003).
7. Lande D.: Creation of subject domain models on the basis of monitoring of network information resources. In:1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, pp. 25–27 (2017). [http://colins.in.ua/wp-content/uploads/2017/04/Lande.pdf]
8. Lande, D., Dmytrenko, O., Radziievska, O.: 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. pp. 132-145. (2019). ISSN 1613-0073 [http://ceur-ws.org/Vol-2577/paper11.pdf]
9. Mcnamee, P., & Mayfield, J. Character n-gram tokenization for European language text retrieval. Information retrieval, 7(1-2), 73-97. (2004).
10. Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M.: Stemming and lemmatization in the clustering of finnish text documents. In Proceedings of the thirteenth ACM international conference on Information and knowledge management, 625-633. (2004).
11. Xu, J., & Croft, W. B.: Corpus-based stemming using cooccurrence of word variants. ACM Transactions on Information Systems (TOIS), 16(1), 61-81. (1998). doi: 10.1145/267954.267957
12. Jivani, A. G.: A comparative study of stemming algorithms. Int. J. Comp. Tech. Appl, 2(6), 1930-1938. (2011).
13. Porter, M. F.: An algorithm for suffix stripping. Program 14(3), 130-137 (1980). doi: 10.1108/eb046814
14. Porter, M. F. Snowball: A language for stemming algorithms. (2001).
15. Ramos, J.: Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning. vol. 242, pp. 133-142. (2003).
16. Lande, D. V., Snarskii, A. A., Yagunova, E. V., & Pronoza, E. V.: 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, pp. 209-215 (2013). doi: 10.1109/MICAI.2013.33
17. Rajaraman, A., & Ullman, J. D. Mining of massive datasets. Cambridge University Press (2011).
18. Beel, J., Gipp, B., Langer S., & Breitinger C.: Paper recommender systems: a literaturesurvey. In:International Journal on Digital Libraries 17(4), 305–338 (2016).
19. Jones, K.S.: A statistical interpretation of term specificity and its application inretrieval. In:Journal of documentation (2004).
20. Lande, D.V., Dmytrenko, O.O., & Snarskii A.A.: Transformation texts into the complex network with applying visibility 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. pp. 95-106. (2018). [http://ceur-ws.org/Vol-2318/paper8.pdf]
21. Luque, B., Lacasa, L., Ballesteros, F., & Luque, J.: Horizontal visibility graphs: Exact results for random time series. Physical Review E 80(4) (2009). doi: 10.1103/PhysRevE.80.046103.
22. Lande, D.V.: Building of networks of natural hierarchies of terms based on analysis of texts corpora. arXiv preprint arXiv:1405.6068 (2014).
23. Kleinberg, J. M.: Authoritative sources in a hyperlinked environment. In Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms, pp. 668-677. Society for Industrial and Applied Mathematics. (1998).
24. Langville, A.N., & Meyer, C.D.: Google's PageRank and beyond: The science of search engine rankings. Princeton University Press (2011).
25. Haveliwala, T. H.: Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. IEEE transactions on knowledge and data engineering, 15(4), 784-796. (2003).
26. Cherven, K.: Network graph analysis and visualization with Gephi. Packt Publishing Ltd (2013).
27. Kleinberg J.M.: Navigation in a small world. In: Nature 406(6798), p. 845 (2000). doi: 10.1038/35022643.