Determining the Directions of Links in Undirected Networks of Terms

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

[1] Institute for Information Recording of NAS of Ukraine, Kyiv, Ukraine
[2] National Technical University "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
[3] Scientific Research Institute for Informatics and Law of National Academy of Legal Sciences of Ukraine, Kyiv, Ukraine
dwlande@gmail.com dmytrenko.o@gmail.com radeoksa@gmail.com

Abstract. This paper examines and analyzes approaches for constructing network of terms as an ontological subject domain model. In particular, new approaches and rules for determining the syntax and semantic links between terms in the text and the directions of these links between nodes in undirected networks of terms constructed from terms of a thematic text corpus, are proposed and researched. Also, one of the methods for creating terminological ontologies – the algorithm for building the thematic networks of natural hierarchies of terms based on analysis of texts corpora – is considered and used to build a directed network of words and phrases (separate unigrams, bigrams and threegrams). The well-known fairy tale "The story of Little Red Riding Hood" is provided as examples to demonstrate an accuracy of the proposed rules. The Python programming language and its separate functions of a specialized add-in - the module NLTK (Natural Language Toolkit open source library) is used to create the software realization of the proposed and considered approaches and methods. Using the software for modelling and visualization of graphs - Gephi, the built directed networks of terms were visualized for better visual perception. 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.

Keywords: Subject Domain, Terminological Ontology, Network of Terms, Horizontal Visibility Graph, Network of Natural Hierarchies of Terms, Syntax and Semantic Links, Undirected Network, Directed Network.

1 Formulation of the Problem

The development of computer technologies and, in particular, the Internet as the source of information resources and a dynamic source of texts, opens new opportunities to develop and apply the improved methods of their research. There are different methods,

methodologies and techniques of computerized text processing and analysis. The modern software is increasingly in need of ready-made solutions to improve its systems.

It should be noted that it is very important to formalize the knowledge of some subject domain while its studying. This process of representing, formal naming and definition of the categories, properties and relations between the concepts, data and entities is known as ontology modeling of the subject domain. A network of terms can be considered as a model of some subject domain. In this network of terms, nodes correspond to the individual words and phrases in the text and the edges to the links between them. The process of ontology creating is usually very complex and resource-intensive, and besides this, it is still an unsolved scientific and practical problem [1]. A separate step in this formalization is to identify the basic objects. In the case of networks of terms building, this step includes creation dictionaries, thesauruses, and subject dictionaries of terms, which based on the text corpus. The task of effective selection of individual terms from the text corpus and automating such selection is still open, important and completely unresolved [2, 3].

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.

The purpose of this work is to propose and present new approaches for determining the directions of links between nodes in undirected networks of terms built from words and phrases (separate unigrams, bigrams and trigrams) of a thematic text corpus.

2 Method for Building Undirected Networks of Terms

There are several approaches for transforming the texts into a network of terms and different ways to interpret nodes and connections [4, 5]. It leads to different kinds of presentation of these networks [6].

In this work, the compactified horizontal visibility graph (CHVG) algorithm for creating terminological ontologies of subject domains for key terms (separate unigrams, bigrams and trigrams) is used.

2.1 Compactified Horizontal Visibility Graph (CHVG) Algorithm

The horizontal visibility graph (HVG) algorithm [7, 8, 9] is a modification of a common visibility algorithm [10].

In the work [11], the next steps are proposed to build undirected networks of terms using the HVG algorithm. 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 built.

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 $t_k (t_i < t_k < t_j)$.

In the third stage, the network that obtained on in the previous steps is compactified: the nodes that correspond to the same terms are combined into a single node. The obtained undirected network of terms is called the compactified horizontal visibility graph (CHVG) (see fig. 1).

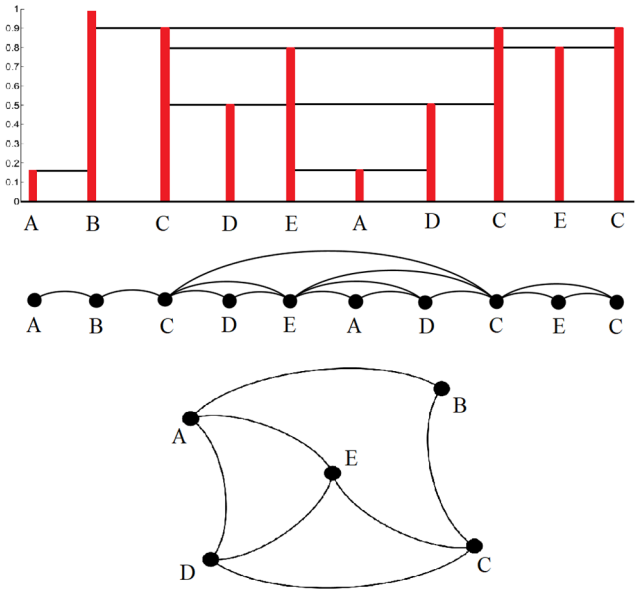

Рис. 1. Stages of a building of the compactified horizontal visibility graph [11].

Thus, the CHVG algorithm allows building an undirected network of terms in case, when the numerical values are assigned to separated words or phrases (separate unigrams, bigrams and trigrams) of a thematic text corpus.

2.2 Text Corpora Pre-processing

Languages we speak and write are made up of several words often derived from one another.

When a language contains words that are derived from another word as their use in the speech changes is called Inflected Language. It is clear to understand that an inflected word(s) will have a common root form.

In this section, we briefly describe the main parts of processing text documents such as tokenization, part-of-speech tagging, lemmatization, stop words removal, stemming process and terms weighting.

Tokenization and lemmatization

For preliminary lexical analysis, breaking text up into its single words (tokens) – *tokenization*, is made.

Lemmatization usually refers to doing things properly with the use of vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. A lemma (plural lemmas or lemmata) is the canonical form, dictionary form, or citation form of a set of words.

For example, "runs", "running", "ran" are all forms of the word "run", therefore "run" is the lemma of all these words. Because lemmatization returns an actual word of the language, it is used where it is necessary to get valid words.

In this work, “WordNet Lemmatizer” provided by Python NLTK was used to lemmatize the tokens. “WordNet Lemmatizer” uses the WordNet Database to lookup lemmas of words.

Tokenization and lemmatization are usually the initial stages of word processing because they allow you to work with a word as a single entity, knowing its context [12].

Part-of-Speech Tagging

POS tagging is one of the first steps in computer text analysis.

Before lemmatization, it is necessary to provide the context in which you want to lemmatize that is the parts-of-speech (POS) [13].

In corpus linguistics, part-of-speech tagging (POS tagging or PoS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context—i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

In general, PoS tagging algorithms are divided into two distinct groups: rule-based and stochastic. E. Brill's tagging method [14], which uses rule-based algorithms, is the first and most widely used method of tagging English-language texts.

“Part of Speech tagging” is one of the more powerful aspects of the NLTK module in the Python programming language. Basically, the goal of a POS tagger is to assign linguistic (mostly grammatical) information to sub-sentential units - tokens.

Stop Words Removal

Also, after the stage of pre-procession of the textual documents and the extraction of key terms in this study, it is proposed to remove stop words that have no semantic strength, that is, informationally unimportant ones, as well as bigrams containing at least one stop word and trigrams that start or end with a stop word. In general, stop words are words that do not contain important significance to be used in Search Queries. Usually, these words are filtered out from search queries because they return a vast amount of unnecessary information. Mostly they are words that are commonly used in the English language such as 'as, the, be, are' etc.

The stop dictionary used in this work was based on different stop dictionaries, which are available at:

https://code.google.com/archive/p/stop-words/downloads/;

http://www.textfixer.com/tutorials/common-english-words.php.

It should be noted, that each programming language will give its list of stop words to use. In this work, the “SnowballStemmer” (stemmer that is realized in Python in NLTK library – Natural Language Toolkit library) was also used to ignore stop words.

Also, the formed stop dictionary was expanded by adding other stop words that were identified by experts within the considered subject domain.

Stemming

After the stages described above, for combining the words that have a common root into a single word it is proposed to carry out the process of stemming. *Stemming* is the process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language [15]. Stemming usually refers to a crude heuristic process that chops off the ends of words and often includes the removal of derivational affixes that are used with a word. So words having the same stem will have a similar meaning. The results of stemming are similar to determining the root of the word, but its algorithms are based on other principles [16]. That is why, after stemming (processing with stemmer), the word may be different from its morphological root.

The goal of both *stemming* and *lemmatization* is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

However, the two processes differ in that stemming most commonly collapses derivationally related words, whereas lemmatization commonly only collapses the different inflectional forms of a lemma.

If confronted with the token "saw", stemming might return just s, whereas lemmatization would attempt to return either "see" or "saw" depending on whether the use of the token was as a verb or a noun.

To avoid the confusion described above, in this work the lemmatization process precedes the stemming process.

Several stemming algorithms can be distinguished in terms of performance, accuracy, and how stemming problems are overcome [17].

The most common algorithm for stemming English, and one that has repeatedly been shown to be empirically very effective, is Porter's algorithm [18, 19]. In this work, the “PorterStemmer” stemmer realized in Python in NLTK (Natural Language Toolkit) library was used. This function is known for its simplicity and speed. As a result of its use, words having the same stem will have a similar meaning.

The pre-processing stages described above allows normalizing the text corpus.

2.3 Weighting and extraction of the key terms

After the pre-processing stages, the weighting and extraction of the key terms are made. To form a time series, the function that reflects the term to number, this study uses the modification of classic statistical weight indicator TF-IDF (from English, TF is Term Frequency, IDF is Inverse Document Frequency) [20, 21] – GTF (Global Term Frequency) [22] as a weight value of terms.

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

3 Rules for Determining the Directions of Links

As was mentioned above, the determination of the directions of links is a complex and open problem of ontology creation. Below, we consider several new approaches for determining the directions of links between nodes in undirected networks of terms built from words and phrases (separate unigrams, bigrams and trigrams) of a thematic text corpus.

Let G be the undirected network of terms that built according to the described above rules: $G:=(V, T)$ where V is the set of nodes, T is the set of the unordered pairs of nodes from the set V that correspond to the causal links between the nodes.

It is supposed that a causal link exists in the direction from the node t_i to the node t_j for $\forall_{i,j}$: $(t_i, t_j) \in T$ if:

1. the numerical value of the node t_i that corresponds to: a) degree [23, 24] b) HITS score [25] c) PageRank score [26]) is higher than the numerical value of the corresponded score of the node t_j;
2. within the sentence, the term to which the node t_i corresponds precedes the term to which the node t_j corresponds;
3. the term to which the node t_i corresponds is shorter than the term to which the node t_j corresponds.

One of the methods for creating terminological ontologies – the algorithm for building the thematic networks of natural hierarchies of terms based on analysis of texts corpora – is used to build the directed network of words and phrases (separate unigrams, bigrams and trigrams) according to the third rule. The work [27] notes that the algorithm for building the networks of natural hierarchies of terms provides for the building of a compactified horizontal visibility graph and the determining of directions of links between the key terms according to the rule: a word is a part of a two-term phrase or a three-term phrase and the two-term phrase is a part of the three-term phrase.

4 Results of the Study of the Proposed Approaches

The proposed approaches for determining the directions of links in undirected networks of terms was tested on the example of the English-language text, namely – the well-known fairy tale “The story of Little Red Riding Hood”.

According to the described above method, the text pre-processing and the extraction of the key terms (separate unigrams, bigrams and trigrams) were made (Table 1, 2 and 3).

Table 1. Top 16 key unigrams and their degree, HITS and PageRank for the text "The story of Little Red Riding Hood".

Unigrams	GTF	Degree	HITS	PageRank
grandmoth	0.065	49	0.444	0.0545
red	0.059	32	0.3099	0.036
hood	0.053	22	0.256	0.0252
ride	0.053	2	0.051	0.0029
wolf	0.031	30	0.301	0.0327
wood	0.025	17	0.204	0.0169
bed	0.019	18	0.191	0.0186
open	0.016	13	0.19	0.0148
beauti	0.016	15	0.155	0.0154
big	0.012	9	0.088	0.0119
cap	0.012	12	0.162	0.0141
cake	0.012	9	0.101	0.0116
cut	0.009	10	0.095	0.0127
strang	0.009	13	0.116	0.0167
ate	0.009	7	0.134	0.0081
huntsman	0.009	11	0.113	0.0148

Table 2. Top 15 key bigrams and their degree, HITS and PageRank for the text "The story of Little Red Riding Hood".

Bigrams	GTF	Degree	HITS	PageRank
ride_hood	0.053	26	0.465	0.0277
red_ride	0.053	28	0.494	0.0303
grandmoth_big	0.009	11	0.177	0.0095
hood_grandmoth	0.006	4	0.107	0.0047
leav_path	0.006	7	0.162	0.0084
grandmoth_live	0.006	6	0.164	0.0071
wolf_bodi	0.006	7	0.160	0.0080
grandmoth_bed	0.006	5	0.04	0.0060
straight_grandmoth	0.006	7	0.133	0.0076
beauti_wood	0.006	6	0.133	0.0070
cake_wine	0.006	8	0.172	0.0084
wood_wolf	0.006	7	0.193	0.0076
press_latch	0.006	8	0.047	0.0064
grandmoth_sick	0.006	5	0.096	0.0058
door_open	0.006	8	0.109	0.0085

Table 3. Top 26 key trigrams and their degree, HITS and PageRank for the text "The story of Little Red Riding Hood".

Trigrams	GTF	Degree	HITS	PageRank
red_ride_hood	0.1429	36	0.67	0.1042
grandmoth_what_big	0.0252	6	0.145	0.0114
press_the_latch	0.0168	6	0.059	0.0111
leav_the_path	0.0168	4	0.153	0.0126
sick_and_weak	0.0168	6	0.182	0.0179
bed_and_pull	0.0168	7	0.162	0.0188
cake_and_wine	0.0168	5	0.160	0.0133
hear_how_beauti	0.0084	2	0.111	0.0076
look_so_strang	0.0084	2	0.03	0.0071
hood_and_ate	0.0084	2	0.111	0.0079
listen_littl_red	0.0084	2	0.129	0.007
bite_he_climb	0.0084	2	0.001	0.0101
obey_her_mother	0.0084	2	0.021	0.0084
lay_the_wolf	0.0084	2	0	0.0105
mind_your_manner	0.0084	2	0.054	0.0068
open_hi_belli	0.0084	2	0.003	0.0096
cake_and_drank	0.0084	2	0.018	0.0090
snore_veri_loudli	0.0084	2	0	0.0105
strang_oh_grandmoth	0.0084	2	0.028	0.0059
bird_are_sing	0.0084	2	0.021	0.0084
bed_fell_asleep	0.0084	2	0	0.0103
ride_hood_enter	0.0084	2	0.114	0.0074
larg_heavi_stone	0.0084	2	0.004	0.0093
woman_wa_snore	0.0084	2	0.111	0.0076
loudli_a_huntsman	0.0084	2	0	0.0105
red_ride_hood	0.0084	2	0	0.1042

The following results were obtained after building the directed network according to the first rule for different measures of network nodes (for the degree – fig. 2; for the HITS – fig. 3; for the PageRank – fig. 4). Using the software for modeling and visualization of graphs – Gephi (https://gephi.org), the built directed networks of terms were visualized for better visual perception.

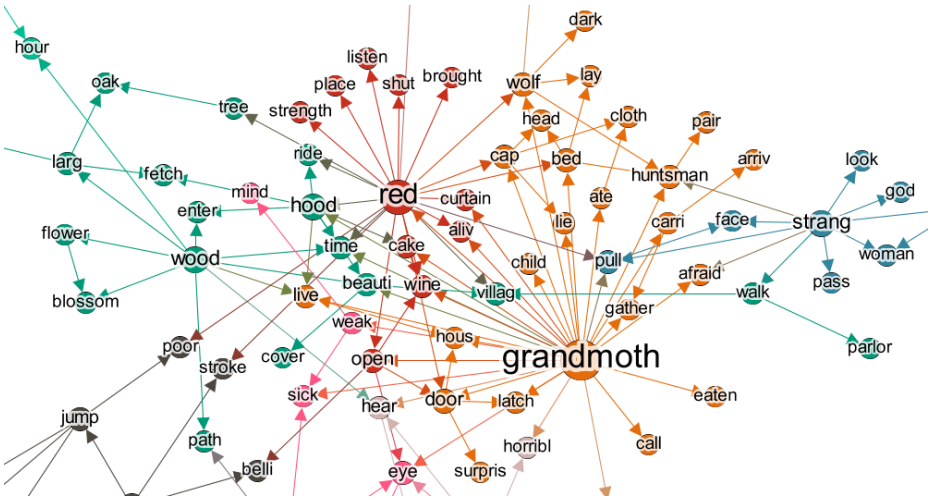

Fig. 2. Fragment of the directed network, which built according to the first rule for node degree.

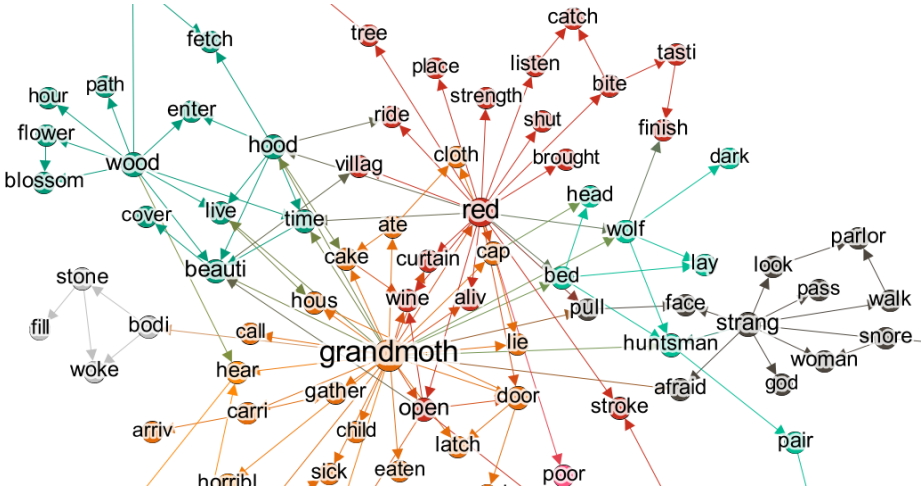

Fig. 3. Fragment of the directed network, which built according to the first rule for HITS.

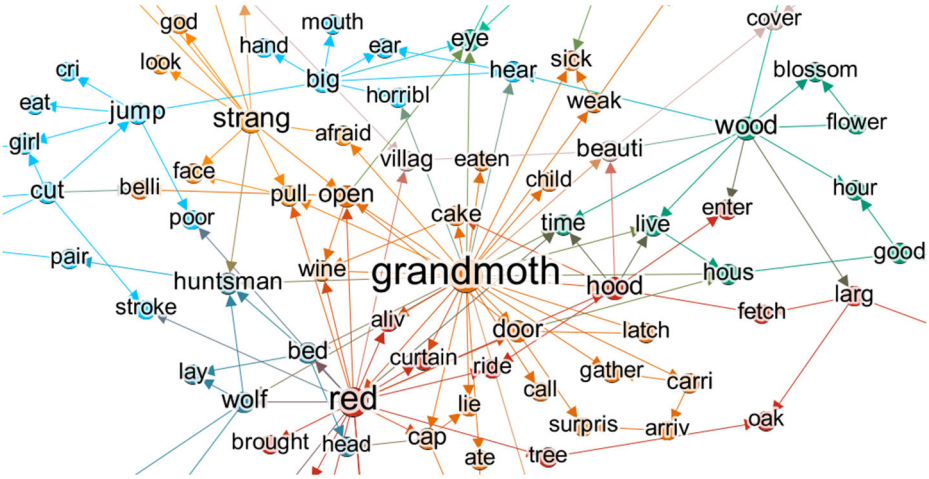

Fig. 4. Fragment of the directed network, which built according to the first rule for PageRank.

Fig. 5 shows the directed network of terms built according to the second rule.

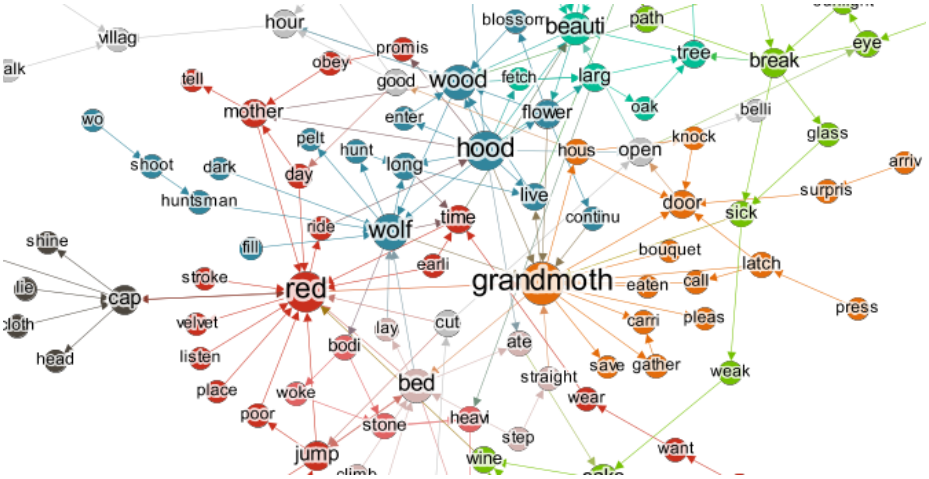

Fig. 5. Fragment of the directed network, which built according to the second rule.

Fig. 6 shows the network of natural hierarchies of terms, which built according to the third rule.

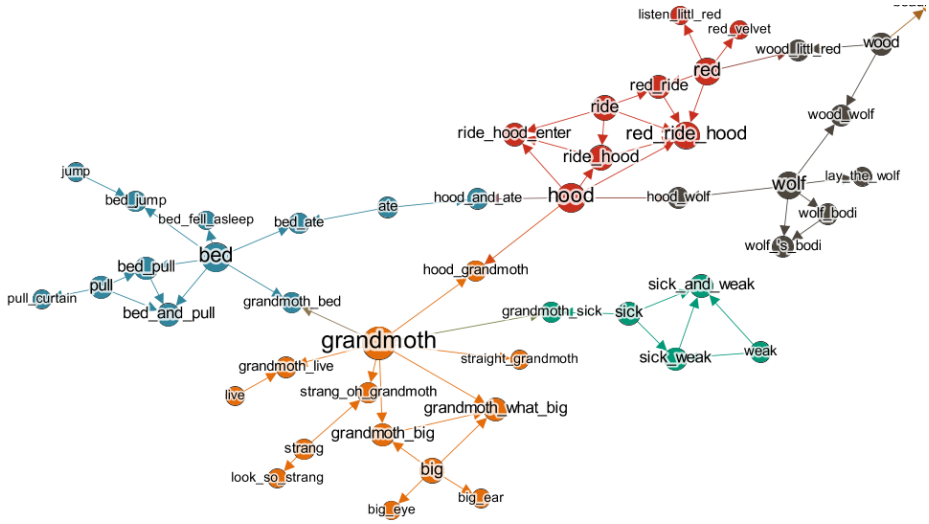

Fig. 6. Fragment of the network of natural hierarchies of terms.

After analyzing the obtained results, it was found that the directed network, which built according to the second rule more precisely reflects the directions of links that exist between the terms in the considered text, than the network, which built according to the first rule. The network of natural hierarchies of terms has its peculiarities and advantages, so it is difficult to compare it with the networks built according to the first two rules. Taking account into the naturalness of links that determined in such a network, we can talk about their syntactic adequacy.

Considering, for example, the directions of links determined for key terms, we can see that according to the first rule, the links between “wolf”-“grandmother”-“red” are as follows (see fig. 2,3,4): for the degree, HITS and PageRank – the “grandmother” influences on the “wolf” and the “red”, and the “red” influences on “wolf”. It does not correspond to the real directions of links that exist in the text in terms of content analysis. While, according to the second rule, the “wolf” influences on the “grandmother” and the “grandmother” influences on the “red”, which corresponds to the content of the considered text.

In comparison with other rules, the rule for determining the directions of links in undirected networks of terms, when within the sentence, the term to which the node t_i corresponds precedes the term to which the node t_j corresponds (where t_j $(t_i, t_j) \in T$) is more informative among the first two rules. It is because the links determined according to this rule more precisely corresponds to the content of the considered text according to experts.

5 Conclusion

After studying the proposed rules for determining the directions of links in undirected networks of terms, it was found that the second rule more precisely reflects the directions of links, which correspond to the content of the considered text according to experts and is more informative. On the example of the English-language text – the well-known fairy tale "The story of Little Red Riding Hood" the undirected network of terms was built. Using the proposed rules for determining the directions of links, the directed networks of terms were obtained from undirected networks of terms. Informative content of network links built according to the second proposed rule is higher among the other two rules according to experts. Taking account into the naturalness of links that determined in the network of natural hierarchies of terms, we can talk about their syntactic adequacy.

The directed networks of words and phrases 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.

As the task of improving the accuracy of determining the directions of links between nodes in undirected networks of words and phrases is actual, then it is planned to continue working in this direction, developing new and modifying existing approaches.

References

1. Lande, D., Snarsky, A.: Approach to Creation of Terminological Ontologies. Design ontology 2(12), pp. 83-91, (2014). (in Russian)
2. Lukashevich, N., Dobrov, B., Chuiko, D.: Selection of Word Combinations for Automatic Word Processing System Dictionary. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue–2008», pp. 339–344. Moscow (2008). (in Russian)
3. Filippovich, Yu., Prokhorov, A.: Semantics of Information Technologies: Experiments of Dictionary-thesaurus Description. Moscow State University of Printing Arts, Moscow (2002). (in Russian)
4. Ferrer-i-Cancho, R., & Solé, R.: The Small World of Human Language. in Proc. of the Royal Society of London, pp. 2261-2265. London (2001).
   doi: 10.1098/rspb.2001.1800.
5. 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. The European Physical Journal B-Condensed Matter and Complex Systems 49(4), 523-529 (2005).
6. Ferrer-i-Cancho, R. F., Solé, R. V., & Köhler, R.: Patterns in syntactic dependency networks. Physical Review E 69(5), (2004).
   doi: 10.1103/PhysRevE.69.051915
7. 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.

8. Gutin, G., Mansour, T., & Severini, S.: A characterization of horizontal visibility graphs and combinatorics on words. Physica A: Statistical Mechanics and its Applications, 390(12), 2421-2428 (2011).
doi: 10.1016/j.physa.2011.02.031.
9. 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, 414, 53-60 (2014).
doi: 10.1016/j.physa.2014.07.002.
10. Lacasa, L., Luque, B., Ballesteros, F., Luque, J., & Nuno, J. C.: From time series to complex networks: The visibility graph. Proceedings of the National Academy of Sciences, 105(13), 4972-4975 (2008).
doi: 10.1073/pnas.0709247105
11. 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).
12. Manning, C. D., Raghavan, P., & Schütze, H.: An Introduction to Information Retrieval. Cambridge University Press, 22–36 (2009).
13. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing, pp. 1–9. Manchester, UK (1994).
14. Brill, E.: A simple rule-based part of speech tagger. In: Proceedings of the third conference on Applied natural language processing (ANLC '92). Association for Computational Linguistics, pp. 152-155. Stroudsburg. PA. USA (1992).
doi:10.3115/974499.974526
15. Jongejan, B., & Dalianis, H.: 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, pp. 145-153. Association for Computational Linguistics, Singapore (2009)
16. Lovins, J. B.: Development of a stemming algorithm. Mech. Translat. & Comp. Linguistics 11(1-2), 22-31 (1968).
17. Baeza-Yates, R., & Ribeiro-Neto, B.: Modern information retrieval. New York: ACM Press, Harlow. England: Addison-Wesle. (2011).
18. Porter, M. F.: An algorithm for suffix stripping. Program 14(3), 130-137 (1980).
doi: 10.1108/eb046814
19. Willett, P.: The Porter stemming algorithm: then and now. Program 40(3), 219-223 (2006).
doi: 10.1108/00330330610681295.
20. Salton, G., & Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management 24(5), 513-523 (1988).
doi:10.1016/0306-4573(88)90021-0
21. Rajaraman, A., & Ullman, J. D. Mining of massive datasets. Cambridge University Press (2011).
22. 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).
23. Bondy, J. A., & Murty, U. S. R.: Graph theory with applications. vol. 290. Macmillan, London (1976).

24. Godsil, C., & Royle, G.: Algebraic Graph Theory. Graduate Texts in Mathematics 207. Springer, New York (2001).
doi: 10.1007/978-1-4613-0163-9
25. Kleinberg, J. M.: Authoritative sources in a hyperlinked environment. In Processing of ACM-SIAM Symposium on Discrete Algorithms, 46(5), pp. 604–632 (1998).
26. Brin, S., & Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, 30(1-7), 107-117 (1998).
doi:10.1016/S0169-7552(98)00110-X
27. Lande, D.V.: Building of networks of natural hierarchies of terms based on analysis of texts corpora. arXiv preprint arXiv:1405.6068 (2014).