UDC 004.912

Formation Networks of Terms for Identifying Semantic Similarity or Difference Degree of Texts in Cybersecurity

Oleh Dmytrenko

National Technical University «Igor Sikorsky Kyiv Polytechnic Institute», Kyiv, Ukraine
Institute for Information Recording of National Academy of Sciences of Ukraine, Kyiv, Ukraine

Abstract
This paper devoted the problem of identifying a semantic similarity degree or difference of text in cybersecurity field. The paper presents a method for comparing text documents based on the formation and comparison of the corresponding semantic networks. The directed weighted network of terms, where the nodes of such networks are key terms of the text, and edges are semantic relationships between these terms in the text are considered as a semantic network. The algorithm for formation semantic networks as one of the types of ontologies is also presented. Formation of network of term includes pre-processing of text data, extraction of key terms, construction of undirected network of terms (using the algorithm of horizontal visibility graph), determining undirected connections between terms, and further determining the directions of connections and their weight values. The Frobenius norm of the difference of matrices corresponding to the semantic networks is considered to compare the semantic networks. An identifying the critically different texts that can have similar keywords but different semantic between them is important to ensure cybersecurity. Also, the proposed approach can be helpful while solving the problem of accumulating text data semantically similar in content. In general, this approach can also be used in systems of automatic information retrieval to determine the degree of similarity or difference in the structure and semantics of texts and identify the sources of information that have a destructive impact on the information space.

Keywords: semantic network, natural language processing, horizontal visibility network, text comparison, computational linguistics, cybersecurity

Introduction

Today, the concept of "Big Data" plays an increasingly important role in the field of cybersecurity. The rapid development of information and telecommunication technologies causes the rapid accumulation of data in various sources – text files, emails, and web pages [1] in various presentation formats. After all, such a process is also associated with the accumulation of a large volume, in particular, of text data. These data can be produced by various sources and have a different nature, including destructive ones. Increasingly, the problem of accumulating text data semantically similar in content arises. Such data is usually informational noise with no additional information. The issue of intentionally entering such data is more complicated. On the other hand, the problem of identifying the critically different texts that can have similar keywords but different semantic between them is also important. Such processes can be malicious in nature and must be detected in order to ensure cybersecurity.

All these problems lead to the need to develop and improve existing technological solutions and create new ones in order to ensure prompt processing and analysis of text information. Taking into account the huge volume of texts, the task of formalizing textual data and presenting them in a form that would be convenient for automatic processing is urgent [2, 3, 4].

The purpose of the paper is to present a method for determining the degree of similarity between text documents, based on the use of directed weighted networks of terms, where the nodes of such networks are key terms of the text, and edges are semantic-semantic relationships between these terms in the text.

1. Formation Networks of Terms

An example of a subject domain model (ontology), which can be represented as a huge array of text data, and which will be convenient for computer processing, is a directed weighted network of terms. Directed Weighted Network of Terms (DWNT) is a semantic model of text representation, where the nodes of such a network are key terms (words and phrases), which are used as the names of concepts in a particular subject area, and the edges is semantic-syntactic connections between these terms. Comparing the DWNTs obtained for different texts, accordingly, allows us to determine the semantic similarity of the respective texts.

Building of network of term is carried out in several stages [3], including pre-processing of text data, extraction of key terms, construction of undirected network of terms (using the algorithm of horizontal visibility graph), ie determining undirected connections between terms, and further determining the directions of connections and their weight values.

For the pre-processing of text data, some of the most common techniques are used, including automatic segmentation into individual sentences and subsequent tokenization of the sentences – segmentation of the input text of sentences into elementary units (tokens) [5]. After tokenization, within each sentence Part-of-Speech tagging (PoS tagging) is doing [6]. PoS tagging consists in assigning each word in the text to a certain part of the language and assigning it a corresponding tag. In addition, in order to obtain canonical, lexical forms of tokens (lemmas), the lemmatization of individual marked tokens is carried out. This step allows to further group different forms of the same word so that they can be analyzed as a single element.

The functions of various Python programming language packages and libraries have been used to computerize word processing, classify tokens, and assign appropriate tags to them. In particular, for the texts presented in Ukrainian and Hebrew, the Pipeline functions of the Stanza library [7] and, accordingly, the English and Hebrew language models were used. Ukrainian and Russian-language texts are processed using the pymorphy2 library [8]. The following link [9] contains a set of predefined tags that the above-mentioned libraries use to match each word in a sentence to a specific part of the language.

For the extracting terms, words related to parts of speech such as noun (NOUN tag), including common names (PROPN tag), adjective (ADJ tag) and conjunction (CCONJ tag) were used.

To build a network of terms, individual words that belong to parts of speech such as nouns (common names with the PROPN tag have been reassigned for convenience) were used. The following templates were used to construct the phrases:

- for bigrams:
  «ADJ_NOUN»;
- for threegrams:
  «NOUN_CCONJ_NOUN»,
  «ADJ_ ADJ _NOUN»;
- for fourgrams:
  «ADJ_NOUN_CCONJ_NOUN»,
  «ADJ_CCONJ_ADJ _NOUN».

Next, the removal of individual stop words (individual articles, prepositions, conjunctions, some verbs, adverbs and pronouns), and which do not have information load is carried out. The list of stop words was formed on the basis of a combination of several stop dictionaries, ones of which for Ukrainian, Russian, English and Hebrew language are available at [10]. And also, each list was expanded with the another available in the Python package – [11]. It is also planned to edit the stop words dictionary by adding and removing from the list of words that have been identified by experts within the research area.

Using keyword and phrase templates, the next step is to form a sequence of terms where more phrases precede the phrases and words that are part of them, with the initial order of occurrence in the sentence being taken into account for single words.

Next stage is to separate the key terms from the text for each formed term of the sequence, the so-called tuple of three elements is built: the first is the term (word or formed according to the presented templates); the next is a tag that is assigned to a word depending on its belonging to a certain part of the language, or a collective tag for the corresponding template; the last element of such a set – the numerical value of GTF (Global Term Frequency) – a global indicator of the importance of the term [2, 4]:

$$GTF = \frac{n_i}{\sum_k n_k}, \qquad (1)$$

where n_i is a number of terms i appearances in the text; $\sum_k n_k$ is a general or global number of formed terms in the whole text.

Taking into account the marking of parts of speech, *GTF* in this case is calculated taking into account the first two elements of the tuple – the term and tag. The number of such identical tuples in the whole sequence, which is normalized to the total number of generated terms, determines the value of the third element of the tuple – *GTF*. Unlike the usual *TF-IDF* statistic, *GTF* allows to more effectively find information-important elements of text when working with a text corpus of a predefined topic, when the information-important term occurs in almost every document in the corpus.

To build an undirected network of terms, as a terminological ontology of a particular subject area, this paper considers and applies an approach to building networks based on time series – Horizontal Visibility Graph algorithm (HVG). The Horizontal Visibility Graph Algorithm (HVG) [12], in turn, is an extension of the standard Visibility Graph Algorithm (VG) [13]. Horizontal visibility graphs are constructed within each individual sentence, where each term corresponds to a statistical estimate *GTF* (Global Term Frequency) – a global indicator of the importance of the term.

An undirected network of terms using the Horizontal Visibility Graph Algorithm is built in two stages [14]. The first step is to mark on the horizontal axis a sequence of nodes ti, each of which corresponds to the terms in the order in which they occur in the text; and the weighted values numerical estimates xi that corresponded to *GTF* and 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. It is considered, 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,

$$x_k < min\,(x_i, x_j), \qquad (2)$$

for all t_k (t_i < t_k < t_j), where $i<k<j$ are the nodes of graph. The obtained undirected network of terms is called the horizontal visibility graph (HVG) (see fig. 1). Therefore, the considered HVG algorithm makes it possible to construct an undirected network structure from time series on the basis of texts in the case when numerical weight values (GTF in our case) are assigned to an individual words or phrases. If a priori there is an undirected connection between the respective nodes in the horizontal visibility graph, the directions of links in an undirected network of terms are established on the principle of entering a shorter term into a term, which is it's an extension. The direction of all other unlinked links is established from left to right (empirical rule).

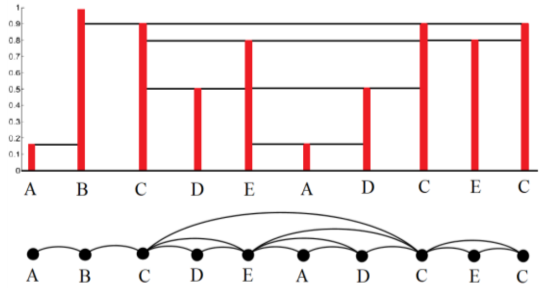

Figure 1: Example of building the Horizontal Visibility Graph

The weight values of the connections between the nodes in the directed network are determined by the principle proposed in [15]: the vertices of the graph corresponding to the same terms of the previously constructed directed network are combined ("merged"). As a result, the weight values of the connections between the pairs of nodes are determined by the number of same directed connections between these nodes. Since any graph is determined by the adjacency matrix, the task of determining the weight values of the links is reduced to the concatenation of columns and corresponding rows, i.e., a weighted compactification of the horizontal visibility graph [14]. The resulting matrix defines an oriented weighted graph formed of vertices that correspond to unique terms in the text. The weight value of the edge, that connect the vertex *i* with the vertex *j* is determined by the number of occurrences of the term t_i before the term t_j in the text.

The resulting network can be saved in "graphml" and json formats. The open-source software package Gephi designed for network analysis and visualization is used to visualize networks presented in "graphml" format. The json format can be convenient for use in systems for building and visualizing semantic networks. During visualization, only the text of the term (words or phrases) is displayed as node labels, without specifying the part of the speech which was assigned to the term at the stage of PoS tagging.

2. Comparison of semantic networks

When comparing the semantic networks considered above, the generally accepted approach is used. Matrix A, which is the difference of matrices corresponding to these semantic networks, is considered. And a norm of matrix A is evaluated as a measure of divergence. The norm of the matrix reflects the order of magnitude of the matrix elements. In this case, it is recommended to use the Frobenius norm $\|.\|_F$, that is equal to the square root of the sum of squares of all elements of the corresponding matrix:

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}. \qquad (3)$$

Of course, the dimension of the two compared matrices must coincide. In reality, the composition of terms in different semantic matrices differs. Therefore, the compared networks are mutually complementary with terms that are part of their overall composition.

3. Example of approbation of the method

The degree of similarity of the texts was determined by the example of biblical texts, which are well-known and translated into almost all languages (in particular, the authors researched texts in Ukrainian, Russian, English and Hebrew). The text of the sacred book of the Torah, the Pentateuch of Moses, was used to build networks of terms and further research. In particular, the Ukrainian translation by Ivan Ogienko was used [16]. The English version of the text is available at the link [17], the Hebrew is available at [18]. The Russian version of the translation made by Archimandrite Macarius is available at the link [19]. In general, all five books "Genesis", "Exodus", "Leviticus", "Numbers" and "Deuteronomy" were researched.

As a result of processing these texts, the networks of terms as ontological models were obtained. Fig. 2 shows a fragment of the network of terms that corresponds to the fourth book "Numbers".

When processing the Pentateuch of Moses, the specifics of the scriptures were taken into account. For instance, the standard list of stop words was modified at the stage of preliminary processing of texts. As a result, a separate list of exception words that in practice do not refer to stop words was formed and conversely, the list of stop words was supplemented by other words that do not have a semantic load within the researched sacred book. The most frequent synonymous words were researched separately, and as a result, a single definite token was assigned. Also, due to the presence of archaisms in similar sacred texts, some words could be assigned incorrect tags during PoS-tagging. Therefore, this issue requires manual processing.

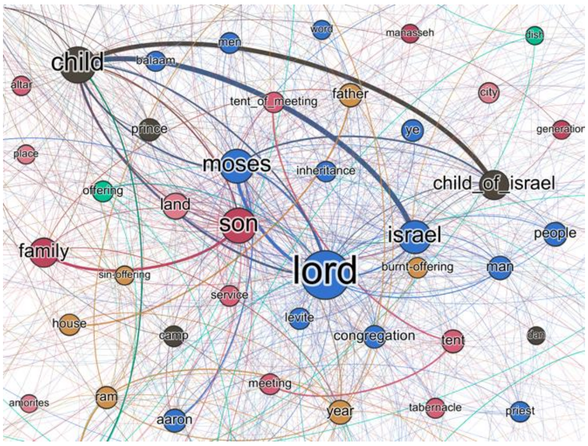

Figure 2: A fragment of the semantic network obtained for the book «Numbers»

Globality in the calculation of GTF was determined within the whole book, or within each individual section, depending on the text for which the network of terms was built i.e. for the entire book or individual section. Therefore, the same terms may have different GTF values within a single section and for the whole text, respectively. These different GTF values affect the build of the horizontal visibility graph.

In order to achieve a slight sparseness of the matrices, the links with a weight equal to 1 were also removed. Next, unconnected nodes with nodes degree equal to 0 were also removed. Such nodes could appear, in particular, after links removing.

All mentioned above affect the topology of networks and lead to the following consequences: the network of terms built for the whole book may contain nodes that do not exist in the network of terms built for an individual section, and vice versa - the network of terms for an individual section may contain nodes that do not exist in the general network built for all text.

Further comparison of the obtained semantic networks built for different texts with applying the Frobenius measure as comparable aproach allows determining the semantic closeness and similarity of the corresponding texts.

The book "Numbers" (the fourth part of the Pentateuch of Moses and the Old Testament) is closest in content to a legal document. This Book contains a census of the adult people of Israel when they were on the Sinai Peninsula and the plains of Moab and regulates the rules of life of these people.

The first part of the book "Genesis" covers 1-10 chapters. They tell of the last days of the people near Sinai.

The second part (chapters 10-22) covers "40 years" in the desert.

The third part (chapters 22-36) describes events in the land of Moab, including Balaam's prophecies about Israel's prosperity.

Semantic networks were also constructed for all chapters of this book (see Fig. 3, 4), which were mutual-semantically compared on the basis of convergence according to Frobenius.

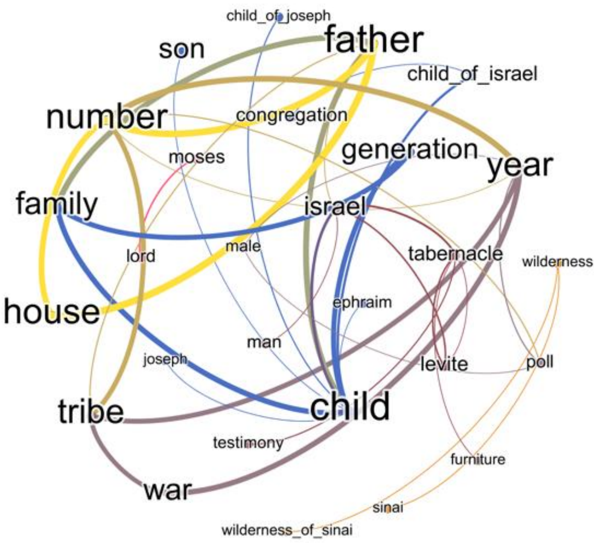

Figure 3: Simplified semantic networks that correspond to 1st sections of the book «Numbers»

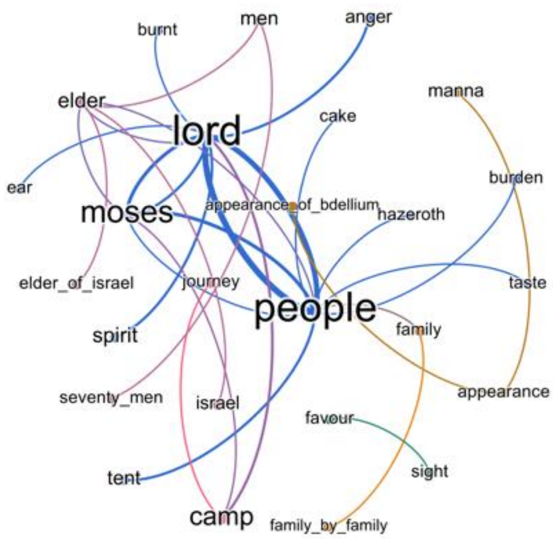

Figure 4: Simplified semantic networks that correspond to 10th sections of the book «Numbers»

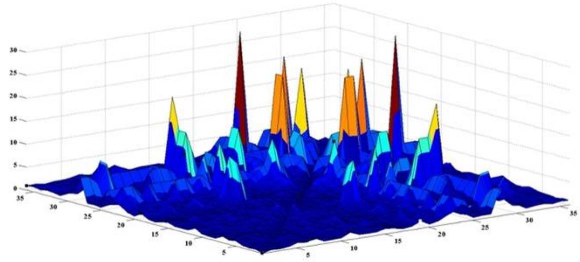

Figure 6: Graph of semantic matrices differences corresponding to separate sections of the book «Numbers»

As can be seen in figure 6, the largest values of the differences correspond to the third part, i.e. sections 22-36. The essence of this anomaly can be found in researchers of the Holy Scriptures. Traditionally, the authorship of the book is attributed to Moses as the author of the Pentateuch. At the same time, it describes the events when Joshua was already chosen as the successor of Moses. Purely narrative fragments in this part of the book are intertwined with legal prescriptions.

That is, the content of the book "Numbers" confirms the network method of the research of text documents to identify structural and terminological differences. The book "Numbers" is the closest in content and structure of the Scriptures to modern legal documents, which suggests that this method can be applied to such documents, in particular, in the exercise of parliamentary control and ensure information security and cybersecurity.

Conclusions

In this paper, the approach for formation networks of terms for identifying a semantic similarity degree or difference of text in cybersecurity field are proposed. The method for comparing text documents based on the formation and comparison of the corresponding semantic networks are presented. The directed weighted network of terms, where the nodes of such networks are key terms of the text, and edges are semantic relationships between these terms in the text are considered as a semantic network. The algorithm for formation semantic networks as one of the types of ontologies is also presented. Formation of network of term includes pre-processing of text data, extraction of key terms, construction of undirected network of terms (using the algorithm of horizontal visibility graph), determining undirected connections between terms, and further determining the directions of connections and their weight values. The Frobenius norm of the difference of matrices corresponding to the semantic networks is considered to compare the semantic networks. An identifying the critically different texts that can have similar keywords but different semantic between them is important to ensure cybersecurity. Also, the proposed approach can be helpful while solving the problem of accumulating text data semantically similar in content. In general, this approach can also be used in systems of automatic information retrieval to determine the degree of similarity or difference in the structure and semantics of texts and identify the sources of information that have a destructive impact on the information space.

References

[1] V. Mayer-Schönberger, K. Cukier, Big data: A revolution that will transform how we live, work, and think, Houghton Mifflin Harcourt, 2013. doi: 10.1093/aje/kwu085

[2] D.V. Lande, O.O. Dmytrenko, O.H. Radziievska, Subject Domain Models of Jurisprudence According to Google Scholar Scientometrics Data, in: Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020). Volume I: Main Conference. Lviv, Ukraine, April 23-24, 2020. CEUR Workshop Proceedings (ceur-ws.org), 2604, 2020, pp. 32-43.

[3] Д.В. Ланде, О.О. Дмитренко, Радзієвська О.Г. "Побудова онтологій в галузі права за даними сервісу Google Scholar." Інформація і право. 28.1 (2019): pp. 74-85.

[4] D.V. Lande, O.O. Dmytrenko, Using Part-of-Speech Tagging for Building Networks of Terms in Legal Sphere. Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021). Volume I: Main Conference Lviv, Ukraine, April 22-23, 2021. CEUR Workshop Proceedings (ceur-ws.org), 2870, 2021, pp. 87-97.

[5] C.D. Manning, P. Raghavan, H. Schütze, "An Introduction to Information Retrieval", Cambridge University Press 39 (2009): 22–36.

[6] B. Santorini, Part-of-speech tagging guidelines for the Penn Treebank Project, Department of Computer and Information Science School of Engineering and Applied Science University of Pennsylvania Philadelphia, PA 19104, 1990.

[7] Stanza – A Python NLP Package for Many Human Languages. URL: https://stanfordnlp.github.io/stanza.

[8] Морфологический анализатор pymorphy2, URL: https://pymorphy2.readthedocs.io/

[9] Universal POS tags, URL: http://universaldependencies.org/docs/u/pos/

[10] Stopwords ISO, URL: https://github.com/stopwords-iso

[11] stop-words 2018.7.23, URL: https://pypi.org/project/stop-words/

[12] B. Luque, L. Lacasa, F. Ballesteros, J. Luque, "Horizontal visibility graphs: Exact results for random time series." Physical Review E 80.4 (2009): 046103. doi: 10.1103/PhysRevE.80.046103

[13] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, J. C. Nuno, From time series to complex networks: The visibility graph, in: Proceedings of the National Academy of Sciences 105.13 (2008): 4972-4975. doi: 10.1073/pnas.0709247105

[14] D.V. Lande, A.A. Snarskii, E.V. Yagunova, E. V. Pronoza, The use of horizontal visibility graphs to identify the words that define the informational structure of a text. in: 12th Mexican International Conference on Artificial Intelligence, 2013, pp. 209-215, doi: 10.1109/MICAI.2013.33.

[15] D.V. Lande, O.O. Dmytrenko.: Methodology for Extracting of Key Words and Phrases and Building Directed Weighted Networks of Terms with Using Part-of-speech Tagging, in: Selected Papers of the XX International Scientific and Practical Conference "Information Technologies and Security" (ITS 2020). CEUR Workshop Proceedings (ceur-ws.org), 2859, 2020, pp 168-177

[16] Біблія (Огієнко), URL: https://uk.wikisource.org/wiki/Біблія_(Огієнко)

[17] Internet Sacred Texts Archive, URL: https://www.sacred-texts.com/index.htm

[18] Hebrew Old Testament, URL: https://www.ccel.org/a/anonymous/hebrewot/home.html

[19] Ветхий Завет (Макарий), URL: https://ru.wikisource.org/wiki/Ветхий_Завет_(Макарий)