UDC: 004.93 DOI: https://doi.org/10.15407/jai2022.01.260

INTELLIGENT TECHNOLOGIES IN INFORMATION RETRIEVAL SYSTEMS

D. Lande[1], A. Soboliev[2], O. Dmytrenko[3]

[1,2] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", c. Kyiv, Ukraine
37, Peremohy av., c. Kyiv, Ukraine, 03056
dwlande@gmail.com

[1,2,3] World Data Center (WDC) for Geoinformatics and Sustainable Development, c. Kyiv, Ukraine
37, Peremohy av., c. Kyiv, Ukraine, 03056

[1,3] Institute for Information Recording of the National Academy of Sciences of Ukraine, c. Kyiv, Ukraine
2, Mykoly Shpaka st., c. Kyiv, 03113

[2] Institute of Special Communications and Information Protection of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", c. Kyiv, Ukraine
37, Peremohy av.., c. Kyiv, Ukraine, 03056

[3] Institute of Artificial Intelligence Problems under MES of Ukraine and NAS of Ukraine, c. Kyiv, Ukraine
40, Academician Glushkov av., c. Kyiv, 03680

[1]http://orcid.org/0000-0003-3945-1178
[2]http://orcid.org/000-0003-4027-042X
[3]http://orcid.org/0000-0001-8501-5313

Abstract. This paper considers the use of modern intelligent technologies in information retrieval systems. A general scheme for the implementation of Internet search engines is presented. The existing and prospective approaches to the intellectualization of individual components of this scheme are presented. An approach to the creation of a system of intelligent agents for information collection is presented. These agents are combined into teams and exchange the results of their work with each other. They form a reliable basis for the information base of search engines, ensure uninterrupted operation of the system in case of failure of individual agents. Methods for the formation of semantic networks corresponding to the texts of individual documents are also considered. These networks are considered as search patterns of documents for information retrieval and detection of duplicates or similar documents. Machine learning methods are used to conduct sentiment analysis. The paper describes an approach that made it possible to make the transition from the use of a naive Bayesian model to a modern machine learning system. The issues of cluster analysis and visualization of search results are also considered.

Keywords: information retrieval, intelligent technologies, information collection agents, sentiment analysis, clustering, natural language processing.

ІНТЕЛЕКТУАЛЬНІ ТЕХНОЛОГІЇ В СИСТЕМАХ ІНФОРМАЦІЙНОГО ПОШУКУ

Д. В. Ланде[1], А. М. Соболєв[2], О. О. Дмитренко[3]

[1,2] Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського", м. Київ, Україна
пр. Перемоги, 37, м. Київ, Україна, 03056
dwlande@gmail.com

[1,2,3] Світовий центр даних (WDC) з геоінформатики та сталого розвитку, м. Київ, Україна
пр. Перемоги, 37, м. Київ, Україна, 03056

[1,3] Інститут проблем реєстрації інформації Національної академії наук України, м. Київ, Україна
вул. Миколи Шпака, 2, м. Київ, 03113

[2] Інститут спеціального зв'язку та захисту інформації Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна
вул. Верхньоключова, 4, м. Київ, Україна, 03056

[3] Інститут проблем штучного інтелекту Міністерства освіти і науки України і Національної академії наук України, м. Київ, Україна
пр. Академіка Глушкова, 40, м. Київ, 03680

[1]http://orcid.org/0000-0003-3945-1178
[2]http://orcid.org/0000-0003-4027-042X
[3]http://orcid.org/0000-0001-8501-5313

Анотація. У цій роботі розглянуто застосування сучасних інтелектуальних технологій в системах інформаційного пошуку. Представлена загальна схема реалізації пошукових систем в Інтернеті і наведені існуючі й перспективні підходи до інтелектуалізації окремих компонентів цієї схеми. Надано підхід до створення системи інтелектуальних агентів збору інформації, які гуртуються в рої та обмінюються між собою результатами роботи, формують надійну основу інформаційної бази пошукових систем, забезпечують безперебійну роботу системи при виході з ладу окремих агентів. Також розглянуто методи формування семантичних мереж, що відповідають текстам окремих документів. Ці мережі серед іншого розглядаються як пошукові образи документів для здійснення інформаційного пошуку і виявлення дублікатів або подібних документів. Методи машинного навчання застосовуються для виявлення тональності текстових повідомлень, сентимент-аналізу. У роботі описано підхід, який дозволив здійснити перехід від застосування наївної байєсовської моделі до сучасної системи машинного навчання. Також розглядаються питання кластерного аналізу і візуалізації результатів пошуку.

Ключові слова: інформаційний пошук, інтелектуальні технології, агенти збору інформації, сентимент-аналіз, кластеризація, прогнозування, обробка природної мови.

Introduction

Methods and means of artificial intelligence, in particular, pattern recognition, machine learning, are increasingly used in all areas of information technology. At the beginning of this century, it was difficult to imagine how widely these methods will be implemented in information retrieval technologies and systems. It was impossible to imagine that these methods can be used in the library or scientific and technical information search practice [1], for example. But the revolution on the Internet, the emergence of Complex Networks concepts [2, 3], Big Data, Text Mining [4], the transition from text search to search for multimedia content have shown that modern search is impossible without intelligent technologies.

Practically, all technological components of modern information retrieval systems currently contain elements of Artificial Intelligence, Machine Learning, Pattern Recognition. This paper will consider components such as the information collection agents' system, syntactic and morphological processing means of natural language, semantic networks formation, sentiment analysis, duplicates or similar documents detection, clustering of search results. These intelligent technologies have been implemented, in particular, in projects such as the system "CyberAggregator" [5], Robusta [6], "X-SCIF" [7] and others.

1.General scheme of information retrieval on the Internet

A modern information retrieval system should ensure the implementation of the following functions, which currently require the use of intelligent technologies:

1) databases formation by connecting to the Internet and collecting according to certain criteria and accounts of information provided in the national coding of certain information resources;
2) setting up a system of agents - automatic scanning modules from websites and social networks;
3) maintaining retrospective databases; creation, indexing of messages in these databases;
4) implementation of full-text search using queries in different languages;
5) initial analysis of text messages stored in system databases;
6) identification of keywords by statistical algorithms in information text provided in different languages recognition of concepts called entities sentiment analysis - determining the emotional tone of information messages;
7) building of semantic networks, information images;
8) detection of duplicates similar in content to information messages;
9) formation of analytical reports, digests, and story chains;
10) data analysis and visualization; visualization of statistical data: by certain sources, number of downloaded messages for the selected period etc. Application of nonlinear dynamics methods for research of thematic information flows (correlation, wavelet, fractal analysis etc.);

11) forecasting events based on the analysis of the dynamics of publications;
12) providing access of many users and software applications to the functional components of the system.

2.Intelligent information collection agents

To enable a simultaneous process of obtaining information from the Internet without the use of third-party services and to control and manage such a system from a single location, the teams (swarms) of intelligent agents are implemented to download data. In doing so, individual agents exchange information with each other, as a result of which they ensure the integrity of the obtained data and distribute the load both on the information collection systems and on the web resources of the information source.

Even when a relatively small modern information retrieval system, the basis for such interaction should be at least 3 servers, which should be located in different data centres and be at great distances from each other, that is cover relatively loosely connected fragments of the Internet (eg, American, Chinese, Russian etc.).

The most common HTTPS protocol can be used to manage and interact between agents. It allows quickly optimize commands for network agents and has the appropriate level of security. As the basis, DBMS NoSQL which can withstand heavy loads, is unpretentious to resources and so on, can be taken to automate and ensure the intelligent process of network agents. This DBMS also allows be easy implemented in the most popular operating systems. At the moment, NoSQL technologies are replacing long-known relational databases. Unlike relational databases, NoSQL databases offer document-oriented data models. So many of these databases run faster, have better scalability, and are easier to use.

Proposed in [8] the agent commands ∇ are a cluster of high availability. When one agent fails in this cluster, its function is taken over by another available agent (fig. 1).

These agent teams are based on the RESTful service architecture, which allows effectively configure and control their work. Also, their interaction is based on system messages such as Heartbeat, which allow effectively investigate the life cycle of agents and in case of any of them failure to quickly identify the problem without losing the data that it has received.

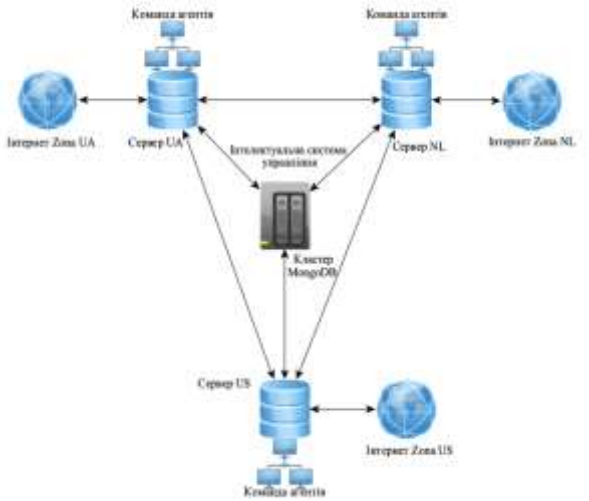

Fig. 1 Scheme of distributed information retrieval based on 3 servers

MongoDB DBMS is used to coordinate the work of the agent system [9]. MongoDB is one of the NoSQL-DBMS, which implements a new approach to databases building. MongoDB has no tables, schemas, SQL queries, foreign keys and many other things that are inherent in object-relational databases. MongoDB is a document-oriented database management system. The main reason for the relational model refusal is the simplification of horizontal scaling and high performance. These advantages determine the possibility of its application for multi-agent systems.

Also, the intelligent control system for network agents works on the following principle: if one of the agents fails, the other is automatically turned on, and in doing so, it is notified of the agent failure. Thus, the information obtaining process from social networks continues to work without stopping due to the intelligent control system of both individual agents and entire swarms. In case the information is changed for a country, the agents while interacting with each other will reflect these changes and keep all copies of received data.

In this case, the general logic of the agent cluster is created at the level of software protocols and allows:

1.To manage all network agents with one intelligent module.

2.To add and improve software and hardware resources without system downtime and large-scale architectural transformations.

3.To ensure uninterrupted operation of the system in case of failure of one or two agents.

4.To synchronize data between agents.

5.To distribute requests to agents effectively.

In fact, the main task of the agent cluster is to eliminate downtime, to collect information that is not available from certain segments of the network. It ensures a completeness and integrity of the information, reporting on the work of individual agents, in particular, determining how much agents collects themselves and how much they takes from other agents. Such an information system can work for a long time without human intervention, only taking into account its information needs, which are entered into a distributed database through a special software interface.

3.Semantic networks for finding and determining the similarity of texts

To determine the degree of similarity of text documents, semantic networks are used. Semantic networks are weighted networks of terms, where the nodes are single words or phrases co-worded from the text, and the edges are semantic-syntactic connections between these formed terms.

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

The building of networks of terms is carried out in several stages [10], including pre-processing of text data, separation of key terms, application of the horizontal visibility graph algorithm to establish undirected links between terms, as well as further establishment of link directions 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 text – segmentation of the input text into elementary units (tokens, tokens). Within each sentence after tokenization, part-of-speech tagging (**PoS tagging**) is marked [11], which consists in assigning a word in the text to a certain part of the language or speech and assigning it a corresponding tag. PoS tagging allows separate words or tokens that can have multiple tags. This step allows further group different forms of the same word so that they can be analysed as a single element.

The functions of the Stanza package of the Python programming language were used for computerized processing of texts presented in the Ukrainian language and classification of tokens by parts of speech and assigning them corresponding tags. The functions of the Stanza package of the Python programming language were used for computerized processing of texts presented in the Ukrainian language and classification of tokens by parts of speech and assigning them corresponding tags. Link [12] contains a set of predefined tags that the Stanza package uses to match each word in a sentence to a specific part of the language. For our purposes, we used words that refer to parts of speech such as nouns (NOUN tag), including common names (PROPN tag), adjectives (ADJ tags), and conjunctions (CCONJ tags).

To build the network, individual words were used that belong to such parts of the language as the noun (common names with the PROPN tag have been renamed the NOUN for convenience). Some adjectives were removed. The following templates were used to build 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 single stop words (individual articles, prepositions, conjunctions, some verbs, adverbs and pronouns), and which do not carry any information load is carried out. The list of Ukrainian stop words was formed on the basis of a combination of several stop dictionaries, one of which is available at [13], and the other is available in the Python package. 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 [10, 14]:

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

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 a 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) [15], in turn, is an extension of the standard Visibility Graph Algorithm (VG) [16]. 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 [17]. The first step is to mark on the horizontal axis a sequence of nodes it, 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,

$$x_k < \min(x_i; x_j)$$

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. 2).

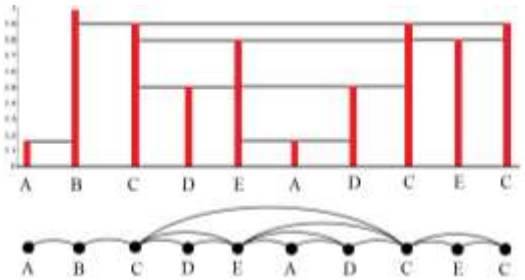

Fig. 2. Stages of building the Horizontal Visibility Graph

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. Directions of links between nodes in the undirected network

of terms are determined 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 defined by t_j [18]. The direction of all other unlinked links is established from left to right.

In work [19], 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.

Next, we 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 [17].

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.

In fig. 3, as an example, a fragment of the building semantic network for the «Genesis» that is the first book of the Pentateuch is considered.

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

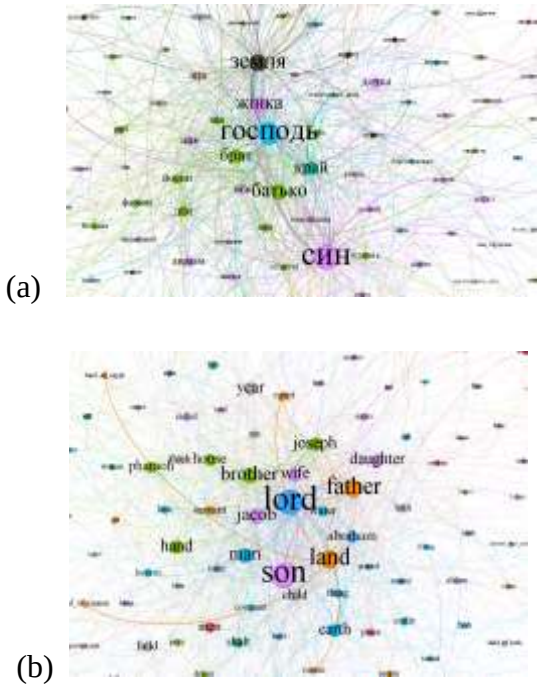

Fig. 3. A fragment of the semantic network obtained for the book "Genesis"
(a – Ukrainian translation, b – English translation)

In general, the proposed method for building semantic networks as one of the types of terminology ontologies obtained on the basis of a thematic corpus of text documents can be used in automated legal information abstraction systems to generate concise information-rich reports based on short annotations or digests. Also, the proposed method can be used in the process of processing information requests during information retrieval, allowing to determine the degree of similarity and semantic closeness of texts to further determine the relevance of the document to the information needs of Internet users and users of information retrieval systems.

The proposed method has two significant advantages over traditional search models. First, the search is not based on an artificially generated query, but on the basis of the text of the primary document that provided by the user (this document can be obtained using traditional search engines), which is almost impossible to apply in other search models. Secondly, all the resulting documents are more pertinent, ie more responsive to the information

needs of the user, which is unattainable in other models.

4.Sentiment analysis

Automated text tone analysis using methods and algorithms based on rules and dictionaries, graph-theoretic models, machine learning models is used to assess the emotional tonality of messages that determined by system requests.

Today, sentiment analysis methods based on the use of machine learning models are the most popular due to their ability to adapt to the recognition of texts belonging to specific discourses. However, the quality of trained models, which is expressed in terms of a recognition accuracy and precision, largely depends on a number of factors, including the method of forming a training corpus of texts. Manual formation of a training data set is a rather time-consuming and costly procedure that is difficult to perform for large thematic texts.

In this regard, a combined approach consisting in the use of simple methods of sentiment analysis (naive Bayesian algorithm) to form a training data set with high recognition accuracy, which is then used to train machine learning models is implemented. To form a text database for machine learning, the documents having the extreme weight values of emotional estimates, which calculated applying the Bayesian algorithm were automatically selected. Thereby, this approach reduces the completeness, but provides the highest allowable accuracy that can be provided by the Bayesian algorithm.

To build a machine learning model, the functionality of the FastText library of Python programming language was used. Classification of moods applying FastText is carried out by loading of the saved quantized trained models. Considering the high speed of obtaining a training set and the quality of the obtained models, it can be considered that the use of simple and fast sentiment analysis methods for automatic formation of training text corpora has an advantage over manual formation of training sets. Also, manual formation training set process is a resource-intensive and requires post-processing by experts. The high speed and quality of the proposed combined approach allows to use it for automatic formation of thematic models of recognition of emotional tonality of thematic texts.

5.Cluster analysis

Cluster analysis is the problem of dividing a given sample of objects into subsets, called clusters, so that each cluster consists of similar objects, and objects of different clusters differ significantly. The problem of clustering refers to statistical processing, as well as a wide range of learning tasks without a teacher.

Cluster analysis of network, as a rule, solves the problem of two-criteria optimization, namely:

1.Within each cluster K, the elements (nodes) should be interconnected as much as possible

$$\sum_{\substack{i\in K\\ j\in K}} a_{ij} \to \max .$$

Here, a_{ij} is an estimate of the relationship between elements with and indexes i and j, which are a part of the cluster K.

2.Onnections between any different clusters, for example, K_p and K_q should be minimal:

$$\sum_{\substack{i\in K_p\\ j\in K_q}} a_{ij} \to \min .$$

Often in the general case, the sums for all indices (N is a number of clusters) are often estimated:

$$\sum_{p=1}^{N}\sum_{\substack{i\in K_p\\ j\in K_p}} a_{ij} \to \max , \quad \sum_{p=1}^{N}\sum_{q=1\atop q\neq p}^{N}\sum_{\substack{i\in K_p\\ j\in K_q}} a_{ij} \to \min .$$

The solution of this formal problem is realized with the help of various methods and algorithms. Among them are the methods of hierarchical agglomerative clustering,
K-means, correlation galaxies, modularity etc.

Modern information retrieval systems and search engines can solve clustering problems for such networks.

Networks of terms (language networks of the subject domain corresponding to the query):

1. Networks of persons. In this case, clusters can correspond to groups of interconnected persons. For example,

terrorist groups or political movements, parties.

2. Network of companies. Firms belonging to common industries, co-owners etc. Can be grouped.

One of the modes of operation of modern information retrieval systems is the formation of digests, which consist of the most important documents on various aspects of the research thema. The necessary number of documents containing the terms with the greatest weight which are included in various clusters from a network of terms is used while the digestformation.

Conclusion

This paper presents several intelligent technologies that allow to perform text search in large databases effectively. Obviously, today this field is the most promising in terms of the application of modern methods of artificial intelligence, including pattern recognition and machine learning.

References

1. Lande D.V., Barkova O.V. Elektronna biblioteka jak seredovishhe adaptivnogo agreguvannja informacij // Bibliotechnij visnik. – 2013. – N 2. – C. 12-17.
2. Newman, M.E.J. The structure and function of complex networks. SIAM Review, vol. 45. pp. 167–256.(2003). doi:10.1137/S003614450342480.
3. Snarskij A.A., Landje D.V. Modelirovanie slozhnyh setej: uchebnoe posobie. – Kiev: Inzhiniring, 2015. – 212 s. ISBN 978-966-2344-44-8.
4. Jan Žižka, František Dařena, Arnoš Svoboda. Text Mining with Machine Learning: Principles and Techniques. – CRC Press, 2020. – 366 p. ISBN 978-113-8601-82-6.
5. Lande D., Subach I., Puchkov O., Soboliev A. A Clustering Method for Information Summarization and Modelling a Subject Domain. Information & Security: An International Journal 50, Iss. 1 (2021): 79-86. doi.org/10.11610/isij.5013.
6. Zgurovsky M., Lande D., Boldak A., Yefremov K., Perestyuk M. Linguistic Analysis of Internet Media and Social Network Data in the Problems of Social Transformation Assessment. Cybern Syst. Anal. 57, 228-237 (2021). doi.org/10.1007/s10559-021-00348-8.
7. Dodonov A.G., Landje D.V., Prishhepa V.V., Putjatin V.G. Komp'juternaja konkurentnaja razvedka. – K.:TOV "Inzhiniring", 2021.–354 s. ISBN 978-966-2344-79-0.
8. Sobolєv A.M., Lande D.V. Rozpodileni intelektual'ni agenti dobuvannja kontentu iz social'nih merezh // Materiali naukovo-praktichnoї konferencії "Informacijno-telekomunikacijni sistemi i tehnologiї ta kiberbezpeka: novi vikliki, novi zavdannja". – Kiїv: ISZZI KPI im. Igorja Sikors'kogo, 2021. – C. 274-275.
9. Shennon Brjedshou, Jon Brjezil, Kristina Hodorov. MongoDB: polnoe rukovodstvo. Moshhnaja i masshtabiruemaja sistema upravlenija bazami dannyh. – M.: DMK Press, 2020. – 540 s.
10. Dmytro Lande, Oleh 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).-Vol-2870.-pp. 87-97. ISSN 1613-0073. [http://ceur-ws.org/Vol-2870/paper9.pdf].
11. 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.
12. Universal POS tags. URL:https://universaldependencies.org/docs/u/pos/.
13. Ukrainian-Stopwords. URL:https://github.com/skupriienko/Ukrainian-Stopwords.
14. Lande, D.V., Dmitrenko, O.O., Radzievs'ka, O.G.: Viznachennja naprjamkiv zv'jazkiv u merezhi terminiv. Informacijni tehnologiї ta bezpeka. Materiali XIX Mizhnarodnoї naukovo - praktichnoї konferencії, ITB-2019, C. 103-112. K.: OOO "Inzhiniring" (2019).
15. 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.
16. 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.
17. 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).
18. 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.
19. Dmytro Lande, Oleh Dmytrenko, Creating Directed Weighted Network of Terms Based on Analysis of Text Corpora, 2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC) (Kyiv, 5-9 Oct. 2020). doi.org/10.1109/SAIC51296.2020.9239182.

Література

1. Ланде Д.В., Баркова О.В. Електронна бібліотека як середовище адаптивного агрегування інформації // Бібліотечний вісник. – 2013. – N 2. – С. 12-17.
2. Newman, M.E.J. The structure and function of complex networks. SIAM Review, vol. 45. pp. 167–256.(2003). doi:10.1137/S003614450342480.
3. Снарский А.А., Ландэ Д.В. Моделирование сложных сетей: учебное пособие. – Киев: Инжиниринг, 2015. – 212 с. ISBN 978-966-2344-44-8.
4. Jan Žižka, František Dařena, Arnoš Svoboda. Text Mining with Machine Learning: Principles and Techniques. – CRC Press, 2020. – 366 p. ISBN 978-113-8601-82-6.
5. Lande D., Subach I., Puchkov O., Soboliev A. A Clustering Method for Information Summarization and Modelling a Subject Domain. Information & Security: An International Journal 50, Iss. 1 (2021): 79-86. doi.org/10.11610/isij.5013.
6. Zgurovsky M., Lande D., Boldak A., Yefremov K., Perestyuk M. Linguistic Analysis of Internet Media and Social Network Data in the Problems of Social Transformation Assessment. Cybern Syst. Anal. 57, 228-237, (2021). doi.org/10.1007/s10559-021-00348-8.
7. Додонов А.Г., Ландэ Д.В., Прищепа В.В., Путятин В.Г. Компьютерная конкурентная разведка. – К.: ТОВ "Інжиніринг", 2021. – 354 с. ISBN 978-966-2344-79-0.
8. Соболєв А.М., Ланде Д.В. Розподілені інтелектуальні агенти добування контенту із соціальних мереж // Матеріали науково-практичної конференції "Інформаційно-телекомунікаційні системи і технології та кібербезпека: нові виклики, нові завдання". – Київ: ІСЗЗІ КПІ ім. Ігоря Сікорського, 2021. – С. 274-275.
9. Шеннон Брэдшоу, Йон Брэзил, Кристина Ходоров. MongoDB: полное руководство. Мощная и масштабируемая система управления базами данных. – М.: ДМК Пресс, 2020. – 540 с.
10. Dmytro Lande, Oleh 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). - Vol-2870. - pp 87-97. ISSN 1613-0073. [http://ceur-ws.org/Vol-2870/paper9.pdf].
11. 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.
12. Universal POS tags. URL:https://universaldependencies.org/docs/u/pos/.
13. Ukrainian-Stopwords. URL:https://github.com/skupriienko/Ukrainian-Stopwords.
14. Ланде, Д.В., Дмитренко, О.О., та Радзієвська, О.Г.: Визначення напрямків зв'язків у мережі термінів. Інформаційні технології та безпека. Матеріали XIX Міжнародної науково-практичної конференції, ІТБ-2019, С. 103-112. К.: ООО "Инжиниринг" (2019).
15. 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.
16. 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.
17. 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).
18. 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.
19. Dmytro Lande, Oleh Dmytrenko, Creating Directed Weighted Network of Terms Based on Analysis of Text Corpora, 2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC) (Kyiv, 5-9 Oct. 2020). doi.org/10.1109/SAIC51296.2020.9239182

Received 07.02.22
Accepted 10.03.22