The Electronic Digests Formation and Categorization for Textual Commercial Content

Lyubomyr Chyrun[a], Vasyl Andrunyk[b], Liliya Chyrun[b], Aleksandr Gozhyj[c], Anatolii Vysotskyi[d], Oksana Tereshchuk[e], Nadiya Shykh[f] and Vadim Schuchmann[g]

[a] *Ivan Franko National University of Lviv, University Street, 1, Lviv, 79000, Ukraine*
[b] *Lviv Polytechnic National University, S. Bandera Street, 12, Lviv, 79013, Ukraine*
[c] *Petro Mohyla Black Sea National University, Desantnykiv Street, 68, Mykolayiv, 54000, Ukraine*
[d] *Anat Company, Chervona Kalyna Avenue, 104, Lviv, 79049, Ukraine*
[e] *Hetman Petro Sahaidachnyi National Army Academy, Heroes of Maidan Street, 32, Lviv, 79012, Ukraine*
[f] *Drohobych Ivan Franko State Pedagogical University, Ivan Franko Street, 24,Drohobych, 82100, Ukraine*
[g] *West Ukrainian National University, Lvivska Street, 11, Ternopil, 46004, Ukraine*

Abstract

This article proposes a practical and logistic method of content processing as a stage of the digests formation. The content processing method describes the digests formation and categorisation as one-step of content life cycle and simplifies the information technology of managing commercial textual content. The paper analyses the main problems of available services for processing commercial content. The proposed method allows you to create tools for processing information resources and implement subsystems for managing commercial content.

Keywords 1

Content, intelligent system, information technology, text mining, information flow, content monitoring, automated abstracting, thematic proximity, source text, latent semantic analysis, commercial content, text analysis, content analysis, internet environment, electronic digest, information retrieval, textual content, content search, text mining method, spatial vector model, information system, text data, modern machine learning technology

1. Introduction

The solution to complex problems in any sphere of life requires information support. Meeting information needs is a prerequisite for innovation. However, the difficulty of obtaining information affects the efficiency and quality of decision-making. The Internet environment is a large-scale mass media (mass media) in various information resource content [1-7]. However, the randomness of the emergence and functioning of information resources, the lack of an apparent periodicity of most of them being maintained and updated, and the existing shortcomings in the implementation of effective content searches do not allow using the Internet environment as the complex media. It is customary to consider individual network elements as full-fledged media. Network media is portals with a certain periodicity of updating, electronic versions of printed periodicals, electronic newspapers, and magazines. The Internet environment does not compete with traditional media in many respects. However, according to some criteria, it has advantages due to its technical characteristics. The Internet environment successfully serves as a source and means of distributing content [8-16].

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: Lyubomyr.Chyrun@lnu.edu.ua (L. Chyrun); vasyl.a.andrunyk@lpnu.ua (V. Andrunyk); lchirun21@gmail.com (L. Chyrun); alex.gozhyj@gmail.com (A. Gozhyj); anat1957@gmail.com (A. Vysotskyi); ok.flyud@gmail.com (O. Tereshchuk); shykh.nadiya@gmail.com (N. Shykh); V.shukhmann@st.wunu.edu.ua (V. Schuchmann)
ORCID: 0000-0002-9448-1751 (L. Chyrun); 0000-0003-0697-7384 (V. Andrunyk); 0000-0003-4040-7588 (L. Chyrun); 0000-0002-3517-580X (A. Gozhyj); 0000-0001-9190-7051 (A. Vysotskyi); 0000-0002-6444-0609 (O. Tereshchuk); 0000-0003-0059-7137 (N. Shykh); 0000-0002-1427-3312 (V. Schuchmann)

2. Related works

A feature of the Internet environment is the constant growth in the production and distribution of various types of content and an increase in its volume to the extent that it makes it impossible to process it directly [1]. There are some specific problems associated with the rapid development of information technology (IT). On the one hand, a robust information array as Internet resources for decision-making in various spheres of life of the state, society and an individual or legal entity. On the other hand, there is a lack of content necessary for decision-making through its dynamics, volumes, production rates, sources and lack of structure [17-19]. The coverage/generalisation of extensive dynamic content information flows continuously generating in the media requires qualitatively new approaches [20-24]. The situation of a sharp increase in the rate of commercial textual content and an increase in its volume in the Internet environment led to several problems [1-8, 11-15]:

- A disproportionate increase in information noise due to poorly structured content;
- The appearance of spurious content (obtained as applications);
- Mismatch of formally relevant content (thematically appropriate) to the actual needs of its consumers;
- Repeated duplication of content (for example, publication in various publications).

The exponential increase in the rate of commercial textual content significantly reduces the efficiency of data processing by traditional methods [25-34]. Essential data are duplicated many times on information resources, which increases according to the exponential law. At the beginning of the computer era, automated word processing programs are created that implemented fragmentation, abstracting, annotation, indexing, and other forms of content analysis and synthesis. However, most processes for creating annotations and automatic abstracting are inefficient; the need for scalable methodologies and information systems (IS) remains [35-42]. There are many ways to solve the problem through two directions: automated abstracting and a summary of the content of primary documents. Automated abstracting based on the extraction of document fragments, selecting the most informative phrases and then forming automated abstracting using them [9]. A summary of the source material is based on selecting the most relevant information from the texts using artificial intelligence methods and unique languages and new readers that summarise the primary documents. Using this approach, you can get more instructions containing information that complements the source text [43-51]. Based on a formal presentation of the semantics of the source document, such systems are configured for a high degree of compression, which is necessary, for example, for sending messages to mobile devices. Therefore, the main difference between the abstracting tools is forming a set of excerpts or a summary of the document. All existing intelligent systems of the Text mining class include digests generation, which are integral module [52-61]. One of the basic procedures for integral systems of this class (digests generation) is automated abstracting based on a large number of documents. For the digest, papers are selected in which the trends of the entire input stream are most clearly reflected. Such digests most closely correspond to the user's information needs, at which this input information stream is formed. Based on the abstract, which amounts to an insignificant part of the source text, users can draw a reasonable conclusion about the primary document, having spent much less effort on this than acquainting it [8]. When automated abstracting, the abstract should be 5-30% of the content source. The preparation of annotations contents (digests) from several sources provides an even greater degree of compression. Automated abstracting is reduced to extracting from contents the minimum relevant fragments using analysis of surface-synthetic relations of lexical units in the text [9-10]. Automated abstracting emphasises the selection of characteristic pieces by the phrasal matching method. As a result, blocks of the most remarkable linguistic and statistical relevance are distinguished. Automatic determination of the frequency of use of combinations and individual words in the source content allows you to determine paragraphs and sentences in which the subject of the content is presented most accurately. The creation of the resulting content is carried out by simply connecting the selected fragments. The generated quasi-abstract is a readable text. The abstract quality depends on the processed text genre. The content of a quasi-abstract depends on other features of the content source. Building a high-quality abstract from fragments of the original document without considering semantic laws is practically impossible for compelling content. The basis of the analytical stage of quasi-reference is the procedure for calculating the weighting coefficients for each block of text following characteristics such as the location of this block in the original, the frequency of occurrence in the text, the frequency of use in key phrases, etc. [20].

Text Mining is a set of text processing methods because new knowledge appears [16]. It is an interdisciplinary field where the essential Data Mining technologies are used in combination with techniques from other research areas such as Information Retrieval (IR), Information Extraction (IE), mathematical linguistics, classification, clustering, and ontology creation [17, 62-65]. In each of these areas, they solve their specific applied problems. Still, it is difficult to draw a clear line between Text Mining and another field of research: they all deal with texts, so the general problems and approaches to solving them intersect. The difference lies in the ultimate goal. For example, in IR, the goal is to find documents that at least partially match the search query and, among those found, select those for which the most complete has expired [17, 66-75]. In addition, Text Mining methods aimed at identifying unknown facts and hidden relationships in the analysis of semantic, lexical and statistical features in arrays of texts. However, the algorithms for this use the same. So IE differs from Text Mining. In this area, we consider methods for extracting specific information, structured data, such as people's names, geographical names, book titles using predefined relationships [18]. It is not known which data can be detected in Text Mining. Text Mining methods are effectively used when creating and populating databases. Due to these circumstances, traditional retrieval IS are gradually losing relevance [1, 76-79]. The reason for this lies not so much in the physical volumes of content flows but in their dynamics, that is, the constant systematic updating of content, far from always-obvious regularity. The coverage and generalisation of large dynamic content flow continuously generated in the media require qualitatively new approaches [1-3].

Find the way out can only in automation tools for identifying the essential components in information flows. In recent years, resource-monitoring systems have increasingly used. This promising area is called content monitoring. Its appearance was due to the systematic tracking of trends and processes in the information environment, which is constantly updated. Text monitoring is most often understood as a meaningful analysis of content flows to obtain the necessary quantitative and qualitative slices, which is carried out continuously for a period not determined in advance [1-9]. The most critical component of text monitoring is content analysis [8]. The technology of practical (in-depth) text analysis of text mining is most often used to obtain slices of information flows. Using computing power allows you to identify relationships that can lead to new knowledge. The task of Text mining is to select the key and most important information for the user. It no needs for the user to view a considerable amount of unstructured content. Developed is based on statistical and linguistic analysis and artificial intelligence methods. Text mining technologies are designed to conduct meaningful research, provide navigation, and search in unstructured texts. Using systems of the Tech mining class, users receive new valuable information in the form of knowledge. The technology of deep text analysis historically preceded deep data analysis, the methodology and approached widely used in Text mining methods. Like most cognitive technologies, Text mining is the algorithmic identification of previously unknown relationships and correlations in existing textual data. The exponential growth for information on the Internet is the reason for the ever-increasing difficulty of finding the necessary documents and organising them in the form of repositories structured by content. It is becoming increasingly difficult for the user to find the required information; traditional search mechanisms are ineffective. Therefore, the topic's relevance is caused by an exponential increase in the number of documents, making it impossible to process data by traditional methods without loss of quality.

The purpose of the study is the design and development of the intellectual components of the automatic formation and categorisation of electronic digests.

It is necessary to solve the following tasks to achieve this goal:

1. To conduct a systematic analysis of the subject area.
2. To develop a module for the formation and categorisation of digests.

The object of research is the analysis of the formation processes (processing, rubric) of information flows in the media. The subject of the study is text processing and digest generation algorithms. The scientific novelty of the obtained research results is due to the set of tasks, consists of comprehensive research and systematisation of theoretical and applied problems of getting the necessary qualitative and quantitative information slices in the form of digests. When solving the tasks, the context model of the system for the automatic formation and categorisation of digests of media publications is proposed for the first time.

3. Material and methods

The key elements include summarisation, selection of phenomena, feature extraction, clustering, classification, question answering, thematic indexing, and keyword searching in Text Mining [1-19]. In addition, there will be supplemented the set using support and creation of taxonomies and thesauri in some cases. Alexander Linden, director of Gartner Research, identified four main types of Text Mining IT applications [20-24].

1. Text classification using statistical correlation to build rules for placing documents in conditional categories. For example, they are used to solve the following problems: grouping documents on Internet networks, placing documents in specific folders, selectively distributing news to subscribers.
2. Clustering is based on document attributes using linguistic and mathematical methods without using conditional categories. They are used when abstracting large document arrays, determining interconnected groups of content, simplifying information visualisation, identifying duplicates or content similar documents.
3. Semantic networks or link analysis determine the appearance of descriptors (key phrases) in a document to provide navigation. The visualisation used for this is a critical link in the presentation of unstructured text document schemes. It is used to present the entire array of content and implement the navigation mechanism for the study of documents and their classes.
4. Extracting facts is intended to obtain facts from the text to improve classification, search, and clustering.

There are several more tasks of Text Mining technology, for example, forecasting and finding exceptions (searching for objects with characteristics, stand out from the crowd) [11].

Stemming reduces a word to the base by discarding auxiliary parts such as the ending or suffix. Stemming is used in content retrieval and linguistic morphology. Many search engines use stemmer from merging, combining words in which the forms after stemmer coincide (consider such words synonyms). This algorithm uses the principle of searching the table, which contains all possible variants of words and their forms after stemming. The advantages of this method are the simplicity, speed, and convenience of handling exceptions to language rules. The disadvantages include the fact that the search table must contain all forms of words.

The presentation of the text depends on the task, which determines the ease and effectiveness of data manipulation. The most widely used view is the Vector Space Model (VSM) [14]. Then a vector whose dimensions are given by the number of text parameters describes the text. The values of these parameters are functions of the frequencies with which these parameters appear in the text box. This process is referred to as a bag model of words since the order. And the relationship between words is not considered. Most of the proposed text presentation algorithms are extensions to the Vector Space Model. Some of them are based on phrases. Others believe the semantics of words or the relation between them. In the third is the hierarchical structure of the text is used [18]. The frequency is with which the term appears in the text body clarifies the meaning of this term in a separate document. The frequency is determined in two ways: to emphasise the presence/absence of a term, it varies within [0; 1], or is specified by a mathematical function. Normalisation is performed taking into account the size of the document, taking into account all unique terms. Statistics are collected for both individual words and phrases. Phrases provide more semantic information than single words, for they give a general idea of the context. Its environment characterises a comment. Through the polysemy of most terms, it is necessary to know at least one phrase that contains the word in question to determine its semantic meaning with greater certainty. In table 1, there are formulated the advantages and disadvantages of presenting a document in individual dishes or whole phrases. The use of terms compensates for the shortcomings of the analysis of individual words and vice versa.

Table 1
Advantages and disadvantages of words and phrases in document presentation

Term	Advantages	Disadvantages
Words	Definition of synonyms.	Lack of contextual information. Problems with finding persistent phrases.

Phrases	Availability of developed tools and algorithms The availability of contextual information. Possibility of stable phrases.	The average value in the case of statistical discourse

The task of classifying a document varies depending on previously identified relationships within the content and between content flows. The classification criteria are preliminarily determined before deciding which algorithm to apply. The grouping criteria are the general theme of the work, the author, relevance or degree of interest in the text by regular users. In the case of thematic classification, they focus on nouns that can characterise the topic. The development of automatic learning methods is the main application for implementing this type of classification. Young and Lew [19] compared some training algorithms, proving, for example, that SVM, k nearest neighbours, and linear regression methods work better than neural networks and the Bayesian approach.

A digest is an annotated text built based on the content analysis. Most algorithms for automatically abstracting documents involve three main stages: analysis of the source text, determination of significant fragments (sentences or entire paragraphs) and the formation of a conclusion. The digest is an annotated source of links to the documents underlying it. When forming digests using quasi-referenced methods, it is almost impossible to obtain a coherent text. The combination of abstracts of each of the content will contain redundant incoherent information. However, subject to abstract submission, consisting of a certain number of announcements of incoming documents and divided into units by these documents, the method described above is quite acceptable [8].

Content monitoring is a semantic analysis of content flows to obtain the necessary quantitative and qualitative sections [1-3]. A typical task of content monitoring is constructing diagrams of the dynamics of the appearance of concepts in time [12].

Automated content monitoring technology has several important features [62-65]:

- Use of the key fragment of the publication as a unit for the formation of a text information array;
- Formation of a bank of critical fragments of publications as a combination of two interrelated processes: synthetic and analytical processing and a multi-level procedure for the content analysis of published texts;
- Indexing key fragments of publications using faceted classification.

Clustering. As a result, the search procedure is presented with lists of documents sorted in descending order of compliance with the query. Inevitable inaccuracies in ranking search results, this type of presentation are not always convenient. Then, the clustering of search results is used, which allows you to submit the results in a generalised form, which simplifies the choice of the area corresponding to the information needs of the user [16, 76-83]. In this case, two classes of clustering methods are used - hierarchical or non-hierarchical. In hierarchical clustering (bottom to top or top to bottom), a cluster tree is formed. Non-hierarchical clustering methods provide high-quality clustering due to more complex algorithms. There is a certain threshold function of the quality of clustering for these methods, the maximisation of which is achieved through the distribution of documents between individual clusters.

Topic indexing (proximity). The vocabulary determines the content subject, and the thematic proximity of the terms is characterised by how often these terms are used in documents of the same issue. It does not always mean the obligatory use of these terms in the same documents.

Denote the thematic proximity of the two terms w_i and w_j as $FSR(w_i, w_j)$. Estimates calculation of the thematic proximity of assignments and tasks of the function $FSR(w_i, w_j)$ is performed according to the terms analysis in an array of content that describes topics [1-8, 66-75]. A matrix A is constructed from the output array of content, the rows of which reflect the terms distribution across the text. As an assessment of the thematic proximity of two terms, the scalar product of the corresponding rows of this matrix is used. To calculate the proximity estimates between all pairs of words, it suffices to calculate the matrix A^T. This approach is similar to the classical methods of presenting information based on a vector-spatial model.

Further development of this approach is the use of the so-called latent-semantic analysis. The matrix *A* is used to construct its approximation of *A*, obtained by latent-semantic analysis. The matrix uniquely defines the thematic proximity function of two terms A^T :

$$FSR(w_i, w_j) = A^T[w_i, w_j]. \tag{1}$$

Note that matrix *A* has a dimension *k*, where *k* is the dimension of the theme space chosen at approximation. With this approach, the complexity of calculating the thematic proximity of two terms is x^k , so it does not depend on the number of documents being analysed and the size of the general dictionary.

Table of interrelated concepts. A concept is used as a basis for grouping documents in an information array (not separate terms, but some semantic entities), which can be expressed in a query language theoretically. In the same way as in the case of individual words, the clustering of documents is compared with the clustering of concepts. In contrast, the concept more accurately reflects the thematic properties of documents. It is achieved by complicating the algorithmic part of clustering. The construction of concept relationship tables (VLT) is based on the language tools of the information retrieval system and cluster analysis methods. The semantic meaning of concepts is determined based on information retrieval language [35-45].

The table of conceptual relationships, which is built as a statistical report that reflects the proximity (joint occurrence in documents) of individual concepts from the real world, is a symmetric matrix *A* – $\| a_{ij} \|$, whose elements a_{ij} are the relationship coefficients of the corresponding pairs concepts. The coefficient a_{ij} corresponds to the number of documents in the input information stream, including ideas (terms or phrases are presented in the language of queries corresponding to the concept *i*), and the coefficient a_{ij}, where $(i \neq j)$ is the number of content in the input stream, which simultaneously corresponds to the concepts *i* and *j*.

Qualitative signs are quite adequately expressed in the information retrieval language. This solution is, in most cases, effective and efficient. The cluster analysis algorithm is used to reorganise concepts to identify block sets of the most related terms.

The Boolean search model is a classic and widely used information representation model (based on set theory) and an information search model (based on mathematical logic) [69-76]. The popularity of this model is due to its ease of implementation, which allows indexing and searching in large document arrays. It is popular to combine a Boolean model with an algebraic vector-spatial model of data representation. On the one hand, this provides a quick search using mathematical logic operators, and on the other hand, high-quality ranking of documents based on keyword weights. Within the framework of the Boolean model, documents and queries are presented in the form of a set of morphemic keyword stems (terms). In a Boolean model, a user query is a logical expression in which keywords (query terms) are associated with the logical operators AND, OR, and NOT. By default, various Internet search engines do not explicitly use logical operations but simply list keywords. By default, it is often assumed that all keywords are connected by a logical AND operation. In these cases, only those contents are included in the search results that contain all the keywords of the query simultaneously. In those systems where the space between words is equated to the OP operator, documents that include at least one of the query keywords are included in the search results . When using the Boolean model, the database consists of an index organised as an inverted array. For each term from the database dictionary, there is a list of documents in which this term occurs. The index can also store the value of the occurrence frequency of a given term in each content; it allows you to sort the list in descending order of occurrence frequency. The classical database, which corresponds to the Boolean model, is organised so that you can quickly access the corresponding list of documents for each term. The structure of the inverted array ensures quick modification when new contents are included in the database. Due to these requirements, an inverted array is often implemented as a B-tree.

Latent-semantic analysis, or indexing, is a method for extracting hidden context-sensitive values of terms and the structure of semantic relationships between them by statistical processing of large sets of text data [8]. This method is widely used in the field of search and the problems of information classification. This approach allows you to automatically recognise the content of the shades of words depending on the context of use. It uses the found indicators of thematic proximity of terms, which are then used to calculate estimates of the thematic proximity of documents. The method is widely used in factor analysis, which is to single out the main factors from the space of elementary ones.

Matrix latent semantic analysis. The mathematical apparatus of this method is based on the singular decomposition of matrices. The technique allows revealing hidden semantic relationships when processing large arrays of documents. As initial information, the latent-semantic analysis uses the same matrix as in the vector-spatial model. Elements of this matrix contain values of the frequency of use of individual terms in documents. From matrix analysis, it is known that any rectangular matrix A can be decomposed into a product of three matrices:

$$A = UXV^T \quad (2)$$

The matrices U and V consist of orthonormal columns, and X the diagonal matrix of singular values, the diagonal elements of the unique numbers of the matrix A, that is, the integral square roots of the eigenvalues of the matrix. The most common option is based on using a matrix schedule for singular values. The original matrix is decomposed into a set of orthogonal matrices, a linear combination of a good approximation of the original matrix.

4. Experiments, results and discussion

The system's ultimate goal is the automated compilation of summaries of materials - digests of electronic publications in the media, extracting the most critical content from one or more documents and the generation of concise and information-rich reports on their basis. The system should carry out information monitoring, receive large volumes of data, analyse, systematise data using an automatic rubricator, accumulate information, index material and save it in the database, solve thematic filtering and generate digests in automatic mode. The choice of the final single compromise solution, considering various criteria, is a rather difficult task when planning and making decisions. Therefore, it is advisable to select relevant information in a hierarchical form using the hierarchy analysis method. Content is chosen using an algorithm that processes the source text and selects its most informationally sound parts, fragments. So at the top level of the hierarchy is the goal - the selection of meaningful information. At the second level are criteria clarifying the purpose: the basis of the text, the completeness of the glossary of terms and the number of keywords. At the third level are alternatives of choice - fragments of the source text (Fig. 1).

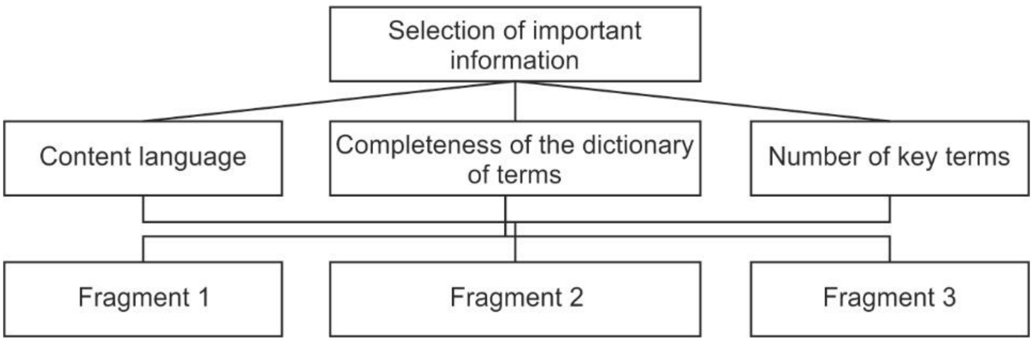

Figure 1: The hierarchical representation of the task

The use case diagram depicts various scenarios of interaction between actors (users) and use cases (use cases); describes the functional aspects of the system. Fig. 2 precedent diagram of a method for automatically forming and categorising digests for electronic media is presented. The article developed by the journalist is processed by the system, in which the statistical indicators of the terms, as a result, are determined. Thematic classification allows you to attribute the article to a specific category. After that, using the Text Mining algorithms, a digest is formed. A class diagram is built to visualise the statistical aspects of the system. Fig. 3 shows a class diagram that describes the system. The Content is part of Analysis, which is part of the Rubricator. The Dictionary is offered as part of the headings and analysis. In Fig. 3b, a state diagram of the system of automatic formation and categorisation of digests for electronic media is presented. It presents a finite state machine with simple states and transitions.

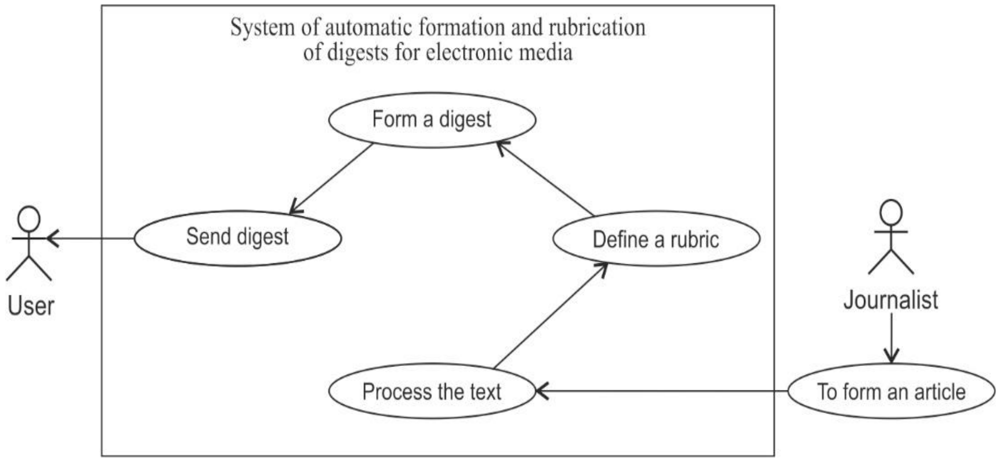

Figure 2: System use case diagram

The purpose of the development is a capable, ready-to-use intelligent system of automatic formation and categorisation of electronic digests. The Web monitoring module allows you to bypass user-specified pages and download updates to Web pages. Special modules that are focused on receiving information of this type read the data received. After receipt and preliminary study, the categorisation module processes all materials. First, the construction and training of the rubricator by an expert is required. The essence of activity is in the expert analysis of educational materials with their classification in one or another rubric; the expert must indicate the degree of the relation of the given text to a particular topic. An activity diagram is a diagram on which the schedule of some activities on its parts is presented (Fig. 4). By activity diagram, we mean a specification of behaviour performed in the form of coordinated sequential and parallel execution of subordinate elements - nested activities and separate actions (actions), interconnected by flows that go from the outputs of one node to the inputs of another. An analogue of activity diagrams is algorithm diagrams.

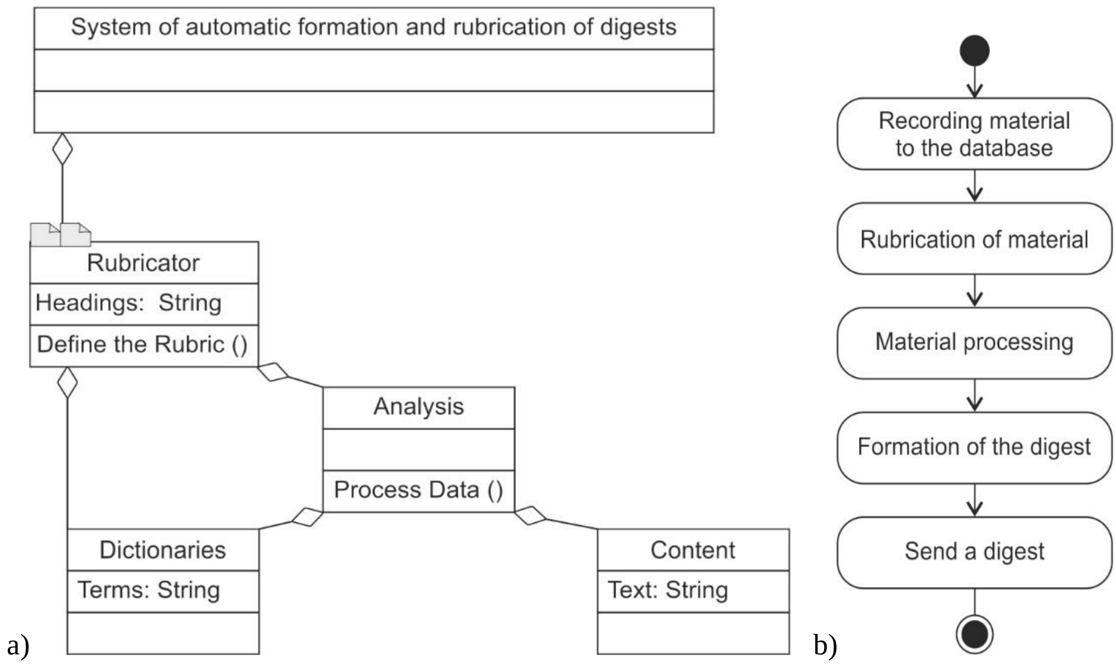

Figure 3: Diagram of (a) classes and (b) states

The sequence diagram shows a message exchange (a method call) between several objects in a particular limited time situation. The time is also steep, directed downward, and arrows with the terms of the operation and parameters (Fig. 5) indicate messages sent from one object to another.

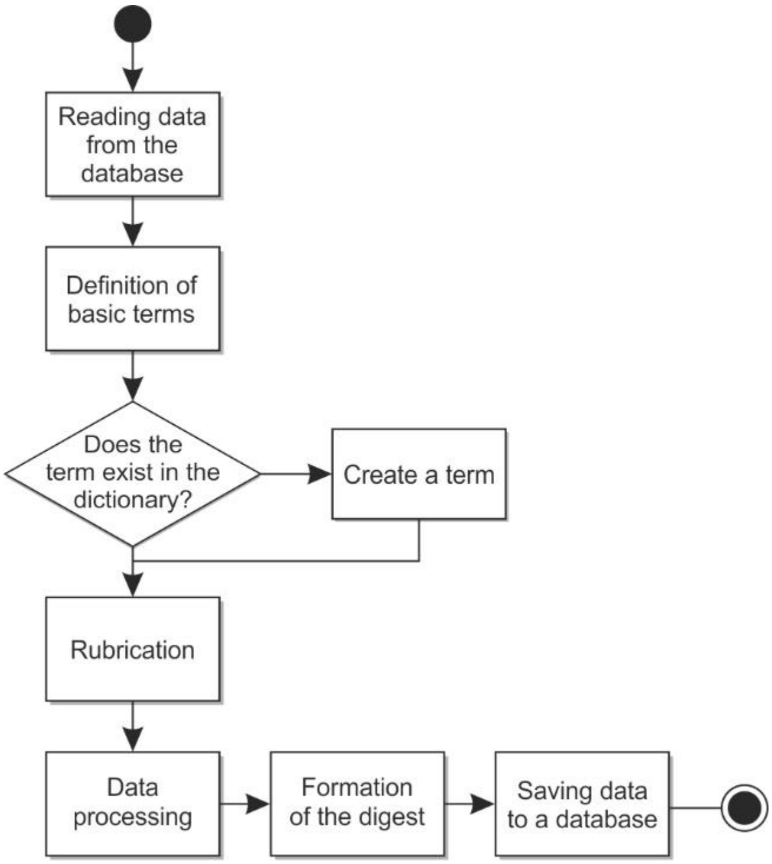

Figure 4: Activity diagram

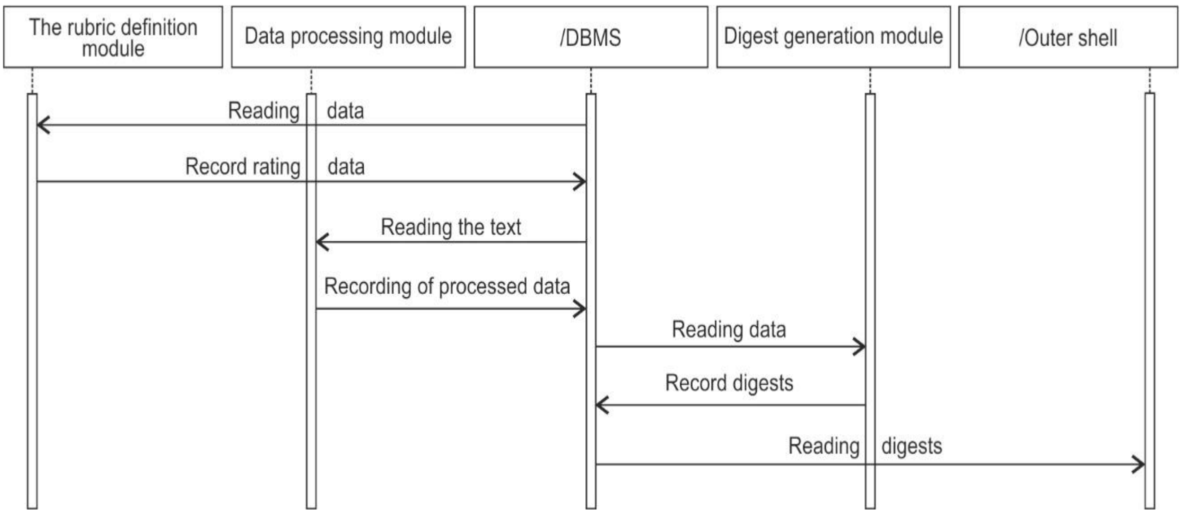

Figure 5: Sequence diagram

The cooperation diagram (Fig. 6, a) is intended to specify the structural aspects of the system's interaction. Cooperation is helpful when modelling design patterns. Based on expert assessments, the categorisation module conducts a semantic and morphological analysis of texts, highlighting the main thematic concepts and analysing the structure of their placement in the text. The systematisation of data in automatic mode is based on the learning outcomes. The source text refers to one or more rubrics with affixing the degree of relationship. After categorisation, the materials should be indexed for all the words of their content. This procedure provides flexible search options based on material attributes and content. The results of the system are presented in the form of thematic digests, which are created automatically. When implementing text processing, it is necessary to consider the existing problems of the formation of dictionaries, the definition of part of speech and the human factor. Firstly, all dictionaries are different and not equivalent to each other. Most often, the task of distinguishing the meaning of words from each other is not difficult. However, in some cases, different definitions of a word can be semantically close (for example, if they are a metaphor or metonymy).

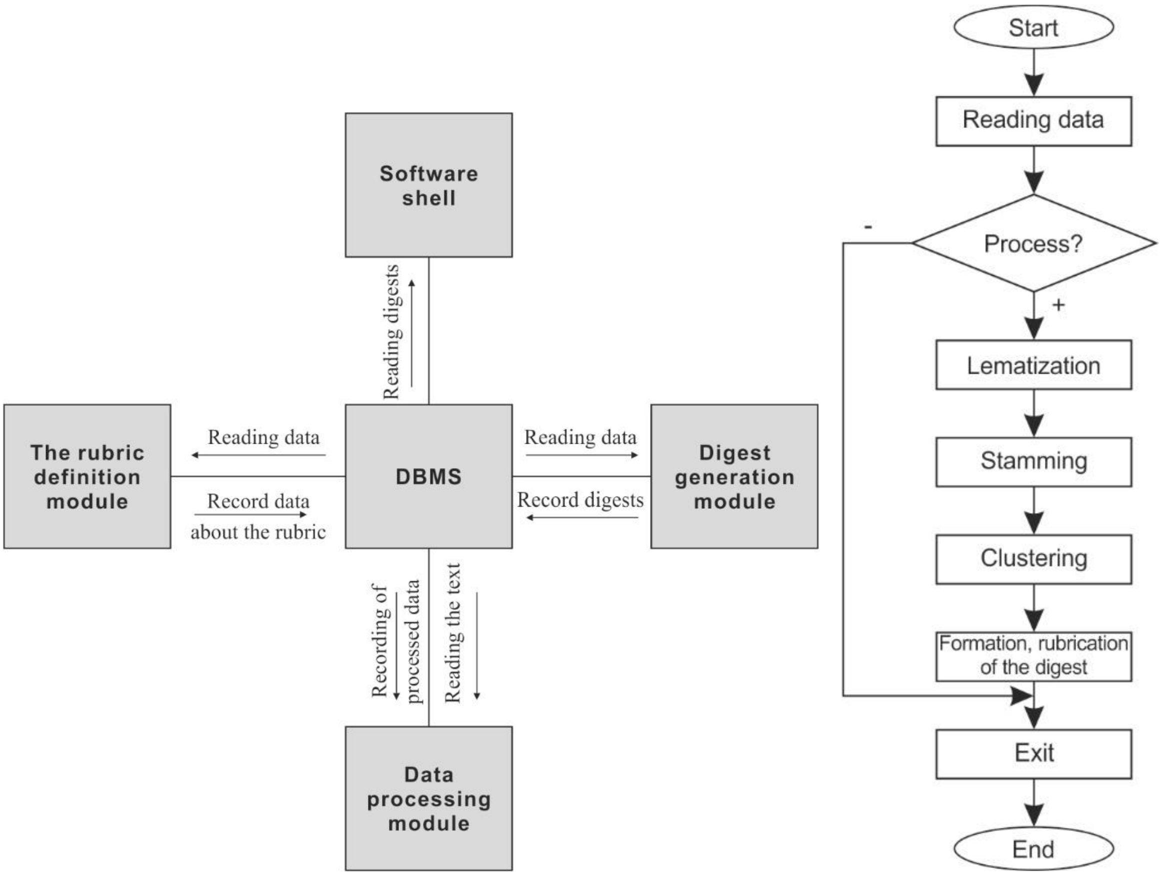

Figure 6: a) Cooperation diagram b) system operation algorithm

In such situations, the separation of meaning in different dictionaries and thesauruses is significant is different. The solution to this problem is the everyday use of the same data source: one standard dictionary. Speaking globally, the results of studies using a more generalised system of separation in meaning are more effective. Secondly, in some languages, determining the part of speech (English Part-of-speech tagging) of a word is very closely related to resolving ambiguity, which these two tasks interfere with each other as a result [7]. There is no consensus, it is worth dividing them into two autonomous components, but the advantage is on the side of those who believe that this is necessary.

The third problem is the human factor. The system is evaluated by comparing the results with the result of the work of experts. In addition, in the case of stylistic writing of articles, the task of constructing meaningful digests and their correct categorisation is performed by an expert.

Digest building. The same strategy of constructing a quasi-abstract is to fix the moments of a profile change, transitioning from low to higher levels and vice versa. The advantage of quasi-reference methods lies in the simplicity of their implementation. However, the selection of text blocks does not consider the relationship between them, which often leads to the formation of inertia essays. Some sentences may appear to be omitted or contain phrases or words that cannot be understood without the preliminary but missing text in the abstract. Attempts to solve this problem boil down to the exclusion of such proposals from abstracts. Less commonly, there are attempts to solve links using linguistic analysis methods.

Unique interfaces are created with the help of which the presence of a significant gap is determined. This approach is not suitable for any mass word processing. As in the case of a quasi-referenced textual content, at the first stage of the formation of the digest, the most significant lexical units are included in the array of output content (input information stream), based on which the dictionary of the system is built. The selection of output documents from the input array of the digest is also carried out, taking into account their weights. Each content weight is determined to consider the sum of the consequences of the individual words included in this content, normalised along the length of the content. The stage of selecting content for the digest consists of such steps as determining the weight of each range, sorting the input document stream by weight, determining the content duplicates of documents according to statistical criteria, rejecting content unsuitable for building digests (invalid types of documents, for example, inspections), and substantial copies (detected by frequency algorithms). The last stage of selecting documents for the formation of the digest is to choose a predetermined number of important content from the array sorted and filtered at the previous locations [8]. The selected contents are submitted in the digest by a predetermined number of significant proposals. In the case of the formation of digests based on the information, it dynamically changes from the Internet; a hypertext representation of the digest is automatically generated, which is considered as an independent document containing links to primary documents on the network. The above procedure provides the formation of a digest that reflects the main trends presented in the source information array. It makes sense to form a fan-shaped multi-aspect digest, reflecting next to the primary trend several other aspects that are ignored in the first type digests. A multi-aspect digest can be built based on technological solutions used in the previous approach when implementing the following algorithm [8].

Stage 1. Build a digest that reflects the primary trend.

Stage 2. Removal from the input information flow of documents corresponding to the trend that was determined in the previous step.

Stage 3. Build a digest that reflects the primary trend of the rest of the information flow.

Stage 4. Combining received digests.

Stage 5. If necessary (based on the required volumes of the resulting digest), the transition to step 2 is performed.

Consider the algorithm of the system of automatic formation and categorisation of electronic digests (Fig. 6, b). The system receives data from the database in the form of an array of articles. Then it checks the user-defined request for processing the text of the article. When there is an article that needs processing, then the lemmatisation of the text of the article is carried out to determine tokens, stemming from dropping stop words and clustering. Such processing as a result, data on the position and weight of word forms will be collected, which will allow you to place the article in a specific category and build a digest.

Logical database schema. Such informational relations in SQL represent all system characteristics:

- jos_content class is information and metadata of articles posted on an information resource;
- digest - information about created digests (identifier, name, digest text, creation date, category identifier)
- rubric class is rubricked information (identifier, rubric name, weight)
- lexeme class is dictionary token table (identifier, token, counter)
- stem class is a table of standard dictionary forms (identifier, topic, counter)
- stopword class is a table of stop dictionary words (identifier, stop word, weight).

The phpMyAdmin admin interface represents the database structure for the intelligent system for the automatic generation and categorisation of digests in Fig. 7.

The central software units that provide the functionality of the system include files:

- default.php is intended to form digest text and performs the function of selecting the short central abstracts from the full text of the article in question;
- stemmer.php is intended for stemming text - cutting off the word endings and suffixes so that the rest, called the stem, is the same for all grammatical forms of the word (in this form, stemmer works only with languages that implement inflexion through affixes)
- lemmatizer.php is intended for lemmatisation of the text - reduction of individual words to standard word forms;
- rubrick.php is an automatic text rubricator and performs the function of identifying the heading for a particular text, using quantitative statistics of the appearance of words in the text obtained using lemmatizer.php;
- content.php is intended to display information on the page.

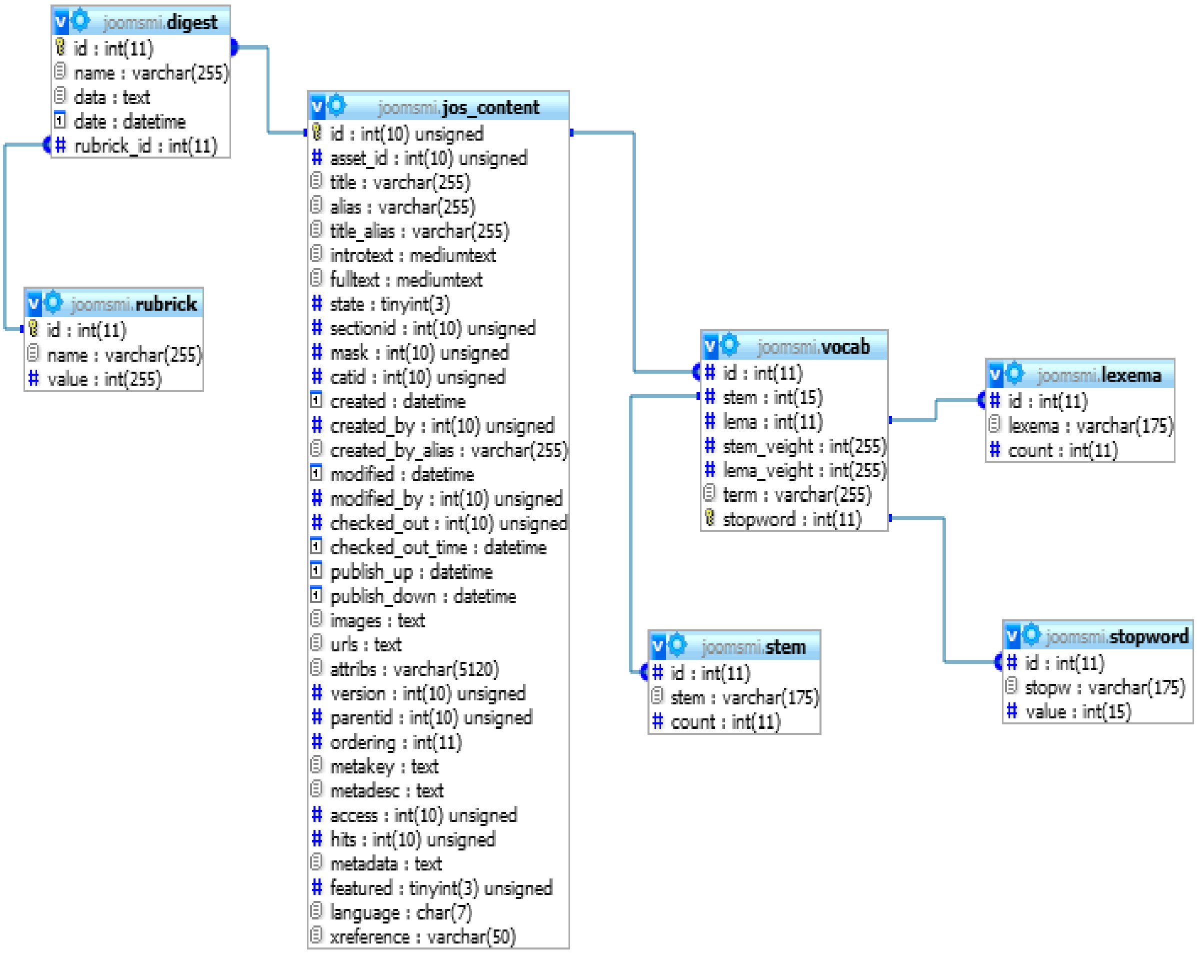

Figure 7: Database structure

5. Conclusions

In the work of information and analytical services, enterprises have to deal with a wide variety of sources of information. These are online newspapers and other Internet resources. In this paper, online media, their disadvantages, advantages, services are considered. Electronic media studies allow us to conclude that the use of human labour in the processes related to the formation and rating of digests is inappropriate. A vital part of this work is developing methods for the construction and categorisation of digests. The experience of implementing the system in various organisations has shown the efficiency and simplicity of adapting the system, thanks to the developed tool for automated generation of digests and their categorisation. A universal data collection module allows you to fully automate the input of electronic information from various sources with its reduction to a single internal format, that is, to minimise the routine work of entering text data. The built-in system for automatically tracking the updating of these pages on information sites on the Internet allows you to automate this part of enterprises' information and analytical services.

6. References

[1] O. Naum, L. Chyrun, O. Kanishcheva, V. Vysotska, Intellectual System Design for Content Formation, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2017, pp. 131-138.

[2] V. Vysotska, L. Chyrun, L. Chyrun, The Commercial Content Digest Formation and Distributional Process, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2016, pp. 186-189.

[3] A. Gozhyj, L. Chyrun, A. Kowalska-Styczen, O. Lozynska, Uniform Method of Operative Content Management in Web Systems, volume 2136 of CEUR Workshop Proceedings, 2018, pp. 62-77.

[4] L. Chyrun, A. Gozhyj, I. Yevseyeva, D. Dosyn, V. Tyhonov, M. Zakharchuk, Web Content Monitoring System Development, volume 2362 of CEUR Workshop Proceedings, 2019, 126-142.

[5] V. Lytvyn, V. Vysotska, Y. Burov, O. Veres, I. Rishnyak, The Contextual Search Method Based on Domain Thesaurus, volume 689 of Advances in Intelligent Systems and Computing, 2018, pp. 310-319.

[6] O. Pavlenko, I. Tymofieieva, Search Query Data Analysis: Challenges and Opportunities, volume Vol-2604 of CEUR workshop proceedings, 2020, pp. 452-461.

[7] N. Antonyuk, M. Medykovskyy, L. Chyrun, M. Dverii, O. Oborska, M. Krylyshyn, A. Vysotsky, N. Tsiura, O. Naum, Online Tourism System Development for Searching and Planning Trips with User's Requirements, volume 1080 of Advances in Intelligent Systems and Computing IV, Springer Nature Switzerland AG, 2020, pp. 831-863.

[8] T. Basyuk, A. Vasyliuk, V. Lytvyn, Mathematical Model of Semantic Search and Search Optimization, volume Vol-2362 of CEUR Workshop Proceedings, 2019, pp. 96-105.

[9] O. Bisikalo, V. Vysotska, Linguistic analysis method of Ukrainian commercial textual content for data mining, volume Vol-2608 of CEUR Workshop Proceedings, 2020, pp. 224-244.

[10] O. Veres, B. Rusyn, A. Sachenko, I. Rishnyak, Choosing the Method of Finding Similar Images in the Reverse Search System, volume 2136 of CEUR Workshop Proceedings, 2018, pp. 99-107.

[11] P. Radiuk, N. Hrypynska, A Framework for Exploring and Modelling Neural Architecture Search Methods, volume Vol-2604 of CEUR workshop proceedings, 2020, pp. 1060-1074.

[12] B. Rusyn, V. Lytvyn, V. Vysotska, M. Emmerich, L. Pohreliuk, The Virtual Library System Design and Development, volume 871 of Advances in Intelligent Systems and Computing, 2019, pp. 328-349.

[13] O. Kliuiev, N. Vnukova, S. Hlibko, N. Brynza, D. Davydenko, Estimation of the Level of Interest and Modeling of the Topic of Innovation Through Search in Google, volume Vol-2604 of CEUR workshop proceedings, 2020, pp. 523-535.

[14] L. Chyrun, I. Kis, V. Vysotska, L. Chyrun, Content monitoring method for cut formation of person psychological state in social scoring, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, pp. 106-112.

[15] D. Hand, N. Adams, Data mining, in: Wiley StatsRef: Statistics Reference Online, 2014, pp. 1-7.

[16] M. W. Berry, J. Kogan, Text mining. Applications and Theory, in: West Sussex, PO19 8SQ, UK: John Wiley & Sons, 2010.

[17] C. C. Aggarwal, C. Zhai, Mining text data. Springer Science & Business Media, 2012.

[18] V. Lytvyn, N. Sharonova, T. Hamon, V. Vysotska, N. Grabar, A. Kowalska-Styczen, Computational linguistics and intelligent systems, volume of Vol-2136 CEUR Workshop Proceedings, 2018.

[19] M. Emmerich, V. Lytvyn, I. Yevseyeva, V. B. Fernandes, D. Dosyn, V. Vysotska, Preface: Modern Machine Learning Technologies and Data Science (MoMLeT&DS-2019), volume Vol-2386 of CEUR Workshop Proceedings, 2019.

[20] V. Lytvyn, N. Sharonova, T. Hamon, O. Cherednichenko, N. Grabar, A. Kowalska-Styczen, V. Vysotska, Preface: Computational Linguistics and Intelligent Systems (COLINS-2019), volume of Vol-2362 CEUR Workshop Proceedings, 2019.

[21] V. Lytvyn, V. Vysotska, T. Hamon, N. Grabar, N. Sharonova, O. Cherednichenko, O. Kanishcheva, Preface: Computational Linguistics and Intelligent Systems (COLINS-2020), volume Vol-2604 of CEUR Workshop Proceedings, 2020.

[22] M. Emmerich, V. Lytvyn, V. Vysotska, V. Basto-Fernandes, V. Lytvynenko, Preface: Modern Machine Learning Technologies and Data Science (MoMLeT+DS 2020), volume Vol-2631 of CEUR Workshop Proceedings, 2020.

[23] V. Lytvyn, V. Vysotska, P. Pukach, Z. Nytrebych, I. Demkiv, A. Senyk, O. Malanchuk, S. Sachenko, R. Kovalchuk, N. Huzyk, Analysis of the developed quantitative method for automatic

attribution of scientific and technical text content written in Ukrainian, volume 6(2-96) of Eastern-European Journal of Enterprise Technologies, 2018, pp. 19-31.
[24] V. Lytvyn, V. Vysotska, P. Pukach, Z. Nytrebych, I. Demkiv, R. Kovalchuk, N. Huzyk, Development of the linguometric method for automatic identification of the author of text content based on statistical analysis of language diversity coefficients, volume 5(2) of Eastern-European Journal of Enterprise Technologies, 2018, pp. 16-28.
[25] V. Vysotska, V.B. Fernandes, V. Lytvyn, M. Emmerich, M. Hrendus, Method for Determining Linguometric Coefficient Dynamics of Ukrainian Text Content Authorship, volume 871 of Advances in Intelligent Systems and Computing, 2019, pp. 132-151.
[26] V. Vysotska, O. Kanishcheva, Y. Hlavcheva, Authorship Identification of the Scientific Text in Ukrainian with Using the Lingvometry Methods, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, pp. 34-38.
[27] V. Lytvyn, V. Vysotska, I. Budz, Y. Pelekh, N. Sokulska, R. Kovalchuk, L. Dzyubyk, O. Tereshchuk, M. Komar, Development of the quantitative method for automated text content authorship attribution based on the statistical analysis of N-grams distribution, volume 6(2-102) of Eastern-European Journal of Enterprise Technologies, 2019, pp. 28-51.
[28] V. Lytvyn, V. Vysotska, Designing architecture of electronic content commerce system. In: Computer Science and Information Technologies, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2015, pp. 115-119.
[29] V. Starko, Semantic Annotation for Ukrainian: Categorization Scheme, Principles, and Tools, volume Vol-2604 of CEUR workshop proceedings, 2020, pp. 239-248.
[30] I. Gruzdo, I. Kyrychenko, G. Tereshchenko, O. Cherednichenko, Applıcatıon of Paragraphs Vectors Model for Semantıc Text Analysıs, volume Vol-2604 of CEUR workshop proceedings, 2020, pp. 283-293.
[31] O. Artemenko, V. Pasichnyk, N. Kunanets, K. Shunevych, Using sentiment text analysis of user reviews in social media for e-tourism mobile recommender systems, volume Vol-2604 of CEUR workshop proceedings, 2020, pp. 259-271.
[32] V. Vasyliuk, Y. Shyika, T. Shestakevych, Information System of Psycholinguistic Text Analysis, volume Vol-2604 of CEUR workshop proceedings, 2020, pp. 178-188.
[33] Y. Yusyn, T. Zabolotnia, Methods of Acceleration of Term Correlation Matrix Calculation in the Island Text Clustering Method, volume 2604 of CEUR workshop proceedings, 2020, 140-150.
[34] Y. Burov, V. Lytvyn, V. Vysotska, I. Shakleina, The Basic Ontology Development Process Automation Based on Text Resources Analysis, in: 15th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 2020, pp. 280-284.
[35] V.-A. Oliinyk, V. Vysotska, Y. Burov, K. Mykich, V. Basto-Fernandes, Propaganda Detection in Text Data Based on NLP and Machine Learning, volume Vol-2631 of CEUR workshop proceedings, 2020, pp. 132-144.
[36] O. Kuropiatnyk, V. Shynkarenko, Text Borrowings Detection System for Natural Language Structured Digital Documents, volume 2604 of CEUR workshop proceedings, 2020, pp. 294-305.
[37] M. Sazhok, V. Robeiko, R. Seliukh, D. Fedoryn, O. Yukhymenko, Written Form Extraction of Spoken Numeric Sequences in Speech-to-Text Conversion for Ukrainian, volume Vol-2604 of CEUR workshop proceedings, 2020, pp. 442-451.
[38] R. Bekesh, L. Chyrun, P. Kravets, A. Demchuk, Y. Matseliukh, T. Batiuk, I. Peleshchak, R. Bigun, I. Maiba, Structural Modeling of Technical Text Analysis and Synthesis Processes, volume Vol-2604 of CEUR workshop proceedings, 2020, pp. 562-589.
[39] N. Shakhovska, O. Basystiuk, K. Shakhovska, Development of the Speech-to-Text Chatbot Interface Based on Google API, volume 2386 of CEUR Workshop Proceedings, 2019, 212-221.
[40] NB. Shakhovska, R.Yu. Noha, Methods and tools for text analysis of publications to study the functioning of scientific schools, volume 47(12) of Journal of Automation and Information Sciences, 2015, pp. 29-43.
[41] T. Batura, A. Bakiyeva, M. Charintseva, A method for automatic text summarisation based on rhetorical analysis and topic modeling, volume 19(1) of International Journal of Computing, 2020, pp. 118-127.

[42] B. Rusyn, V. Vysotska, L. Pohreliuk, Model and architecture for virtual library information system, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, pp. 37-41.
[43] V. Lytvyn, V. Vysotska, B. Rusyn, L. Pohreliuk, P. Berezin, O. Naum Textual Content Categorizing Technology Development Based on Ontology, volume Vol-2386 of CEUR Workshop Proceedings, 2019, pp. 234-254.
[44] V. Lytvyn, V. Vysotska, I. Peleshchak, T. Basyuk, V. Kovalchuk, S. Kubinska, L. Chyrun, B. Rusyn, L. Pohreliuk, T. Salo, Identifying Textual Content Based on Thematic Analysis of Similar Texts in Big Data, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2019, pp. 84-91.
[45] V. Vysotska, V. Lytvyn, V. Kovalchuk, S. Kubinska, M. Dilai, B. Rusyn, L. Pohreliuk, L. Chyrun, S. Chyrun, O. Brodyak, Method of Similar Textual Content Selection Based on Thematic Information Retrieval, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2019, pp. 1-6.
[46] V. Lytvyn, V. Vysotska, P. Pukach, I. Bobyk, D. Uhryn, Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry and glottochronology, volume 4(2-88) of Eastern-European Journal of Enterprise Technologies, 2017, pp. 10-19.
[47] J. Su, A. Sachenko, V. Lytvyn, V. Vysotska, D. Dosyn, Model of Touristic Information Resources Integration According to User Needs, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, pp. 113-116.
[48] V. Vysotska, Linguistic Analysis of Textual Commercial Content for Information Resources Processing, in: Proceedings of the Modern Problems of Radio Engineering, Telecommunications and Computer Science, TCSET, 2016, pp. 709-713.
[49] V. Lytvyn, V. Vysotska, D. Dosyn, O. Lozynska, O. Oborska, Methods of Building Intelligent Decision Support Systems Based on Adaptive Ontology, in: Proceedings of the IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP, 2018, pp. 145-150.
[50] V. Vysotska, V.B. Fernandes, M. Emmerich, Web content support method in electronic business systems, volume Vol-2136 of CEUR Workshop Proceedings, 2018, pp. 20-41.
[51] J. Su, V. Vysotska, A. Sachenko, V. Lytvyn, Y. Burov, Information resources processing using linguistic analysis of textual content, in: Proceedings of the Intelligent Data Acquisition and Advanced Computing Systems Technology and Applications, Romania, 2017, pp. 573-578.
[52] V. Lytvyn, V. Vysotska, D. Dosyn, Y. Burov, Method for ontology content and structure optimisation, provided by a weighted conceptual graph, volume 15(2) of Webology, 2018, 66-85.
[53] V. Lytvyn, V. Vysotska, A. Demchuk, I. Demkiv, O. Ukhanska, V. Hladun, R. Kovalchuk, O. Petruchenko, L. Dzyubyk, N. Sokulska, Design of the architecture of an intelligent system for distributing commercial content in the internet space based on SEO-technologies, neural networks, and Machine Learning, volume 2(2-98) of Eastern-European Journal of Enterprise Technologies, 2019, pp. 15-34.
[54] V. Vysotska, Y. Burov, V. Lytvyn, O. Oleshek, Automated Monitoring of Changes in Web Resources, volume 1020 of Advances in Intelligent Systems and Computing, 2020, pp. 348-363.
[55] Y. Burov, V. Vysotska, P. Kravets, Ontological approach to plot analysis and modeling, volume Vol-2362 of CEUR Workshop Proceedings, 2019, pp. 22-31.
[56] V. Lytvyn, V. Vysotska, P. Pukach, I. Bobyk, D. Uhryn, Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry and glottochronology, volume 4(2-88) of Eastern-European Journal of Enterprise Technologies, 2017, pp. 10-19.
[57] L. Vasyl, V. Victoria, D. Dmytro, H. Roman, R. Zoriana, Application of Sentence Parsing for Determining Keywords in Ukrainian Texts, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2017, pp. 326-331.
[58] V. Andrunyk, L. Chyrun, V. Vysotska, Electronic content commerce system development, in: Proceedings of the 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM, 2015.
[59] J. Chen, D. Dosyn, V. Lytvyn, A. Sachenko, Smart data integration by goal driven ontology learning, volume 529 of Advances in Intelligent Systems and Computing, 2017, pp. 283-292.

[60] D. Dosyn, V. Lytvyn, V. Kovalevych, O. Oborska, R. Holoshchuk, Knowledge discovery as planning development in knowledgebase framework, in: Proceedings of the Modern Problems of Radio Engineering, Telecommunications and Computer Science, Proceedings of the 13th International Conference on TCSET, 2016, pp. 449-451.
[61] D.G. Dosyn, O.L. Ivantyshyn, V.V. Koshovyy, I.M. Romanyshyn, S.O. Soroka, To a question on the mechanism of formation of ionospheric disturbances at groundbased artificial acoustic excitation, in: Proceedings of the International Seminar/Workshop on Direct and Inverse Problems of Electromagnetic and Acoustic Wave Theory, DIPED, 2003, pp. 211-214.
[62] V. Lytvyn, D. Dosyn, M. Emmerich, I. Yevseyeva, Content formation method in the web systems, volume 2136 of CEUR Workshop Proceedings, 2018, pp. 42-61.
[63] M. Davydov, O. Lozynska, Information system for translation into Ukrainian sign language on mobile devices, in: Proceedings of the International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 2017, pp. 48-51.
[64] O.H. Lypak, V. Lytvyn, O. Lozynska, R. Vovnyanka, Y. Bolyubash, A. Rzheuskyi, D. Dosyn, Formation of Efficient Pipeline Operation Procedures Based on Ontological Approach, volume 871 of Advances in Intelligent Systems and Computing, 2019, pp. 571-581.
[65] M. Davydov, O. Lozynska, Mathematical method of translation into ukrainian sign language based on ontologies, volume 871 of Advances in Intelligent Systems and Computing, 2018, pp. 89-100.
[66] V. Savchuk, O. Lozynska, V. Pasichnyk, Architecture of the Subsystem of the Tourist Profile Formation, volume 871 of Advances in Intelligent Systems and Computing, 2019, pp. 561-570.
[67] V. Lytvyn, The similarity metric of scientific papers summaries on the basis of adaptive ontologies, in: Proceedings of the 7th International Conference on Perspective Technologies and Methods in MEMS Design, MEMSTECH, 2011, p. 162.
[68] V.V. Lytvyn, O.I. Tsmots, The process of managerial decision making support within the early warning system, volume 149(11) of Actual Problems of Economics, 2013, pp. 222-229.
[69] P. Kravets, Game methods of construction of adaptive grid areas, in: Proceedings of the The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM, 2003, pp. 513-516.
[70] P. Kravets, Adaptive method of pursuit game problem solution, in: Proceedings of the Modern Problems of Radio Engineering, Telecommunications and Computer Science Proceedings of International Conference, TCSET, 2006, pp. 62-65.
[71] P. Kravets, Game methods of the stochastic boundary problem solution, in: Proceedings of the Perspective Technologies and Methods in MEMS Design, MEMSTECH, 2007, pp. 71-74.
[72] P. Kravets, O. Prodanyuk, Game task of resource allocation, in: Proceedings of the Experience of Designing and Application of CAD Systems in Microelectronics, CADSM, 2009, pp. 437-438.
[73] P. Kravets, Game method for coalitions formation in multi-agent systems, in: Proceedings of the International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 2018, pp. 1-4.
[74] P. Kravets, R. Kyrkalo, Fuzzy logic controller for embedded systems, in: Proceedings of the International Conference on Perspective Technologies and Methods in MEMS Design, MEMSTECH, 2009, pp. 58-59.
[75] P. Kravets, The game method for orthonormal systems construction, in: Proceedings of The Experience of Designing and Application of CAD Systems in Microelectronics, 2007, 296-298.
[76] P. Kravets, The control agent with fuzzy logic, in: Proceedings of the Perspective Technologies and Methods in MEMS Design, MEMSTECH, 2010, pp. 40-41.
[77] P. Zhezhnych, O. Markiv, A linguistic method of web-site content comparison with tourism documentation objects, in: Proceedings of the International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 2017, pp. 340-343.
[78] P. Zhezhnych, O. Markiv, Recognition of tourism documentation fragments from web-page posts, in: Proceedings of the 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering, TCSET, 2018, pp. 948-951.
[79] P. Zhezhnych, O. Markiv, Linguistic comparison quality evaluation of web-site content with tourism documentation objects, volume 689 of Advances in Intelligent Systems and Computing, 2018, pp. 656-667.