Content Formation Method in the Web Systems

Vasyl Lytvyn[1], Dmytro Dosyn[2], Michael Emmerich[3], Iryna Yevseyeva[4]

[1] Lviv Polytechnic National University, Lviv, Ukraine
[2]Karpenko Physico-Mechanical Institute of the NAS, Lviv, Ukraine
[3]Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands
[4]Computing Science at the Faculty of Technology at De Montfort University, Leicester, UK

Vasyl.V.Lytvyn@lpnu.ua

Abstract. In the given article the main problems of Web systems as example electronic content commerce system and functional services of Web content processing are analyzed. In the given article is functional logistic model of content processing as the Web content lifecycle stage in Web systems is proposed. The proposed logistic model gives an opportunity to create an instrument of information resources processing in Web systems and to implement the subsystem of Web content formation, management and support. The model of Web content processing describes the information resources forming in Web systems and automation information technology simplifies the Web content management. In the given article content is forming method as the content lifecycle stage in Web systems is proposed. The method implements the information resources processing in Web systems and automation technology simplifies the Web content formation. In the given article the main problems of e-commerce and functional services of Web content forming are analyzed. The proposed method gives an opportunity to create an instrument of Web resources processing and to implement the module of content forming. The article purpose is to develop a Web content forming method for information resources processing in Web systems. The work relevance is the need to obtain operational/ objectively assess of the competition level in the financial market segment of Web content; assess the competition level and the competitiveness degree in the financial market with Web content distribution. From the systematic approach standpoint to investigate stages of the Web resources processing and optimal lifecycle develop for the Web content formation. The method development of content forming is enables the means of information resources processing and automatic generation of Web content.

Keywords: information resources, Web content, Web resources, content search, content monitoring, content analysis, content lifecycle, Web systems.

1 Introduction

Internet active development promotes a needs growth in production/strategic data and new forms of information services implementation [1-6]. Documented information is an informational product or Web content, if it is prepared in accordance with user

needs and intended to meet them. Web systems development and implementation is one of the e-business development strategic directions. The characteristic feature of such Web systems is the automatic information resources processing to increase Web content sales of permanent user, for potential users active involvement and expanding the target audience boundaries [6-9]. The actual problem in the Web systems design, development, implementation and maintenance is to the research active development in the e-business [10-14]. An important problem is the lack of theoretical justification, standardized methods and software for Web resources processing in such Web systems [15-19]. The important issue is the discrepancy between the known software and methods of Web resources processing and the Web systems construction principles [20-24]. There is no common approach of Web systems creation and standardized methods of Web resources processing in these systems [25-28]. Specialists in designing, implementation and deployment of electronic content commerce systems (ECCS) as example Web systems deals with the information resources processing at various levels. They contribute to the goal to increase sales volumes of content a regular user, the active involvement of potential users and the boundaries expansion of the target audience [10-15]. The special feature of Web systems is as follows [1-2, 10-15]: open – access for all companies and users; global – access from anywhere in the world; unlimited in time – available at any time of the day/week/year; frankness – a low barrier to market entry; direct interaction with the user – reducing the channels of distribution and elimination of intermediaries production; information products and information services testing and implementation; automatic processing requests; automatically track information about users; reducing costs for the business operation; providing more information in online [29-35].

2 The problem of the relationship with important scientific and practical tasks

Urgency of ECCS implementation due to business globalization; increasing needs of content and quick access to the content for the successful conduct of e-business; uneven operation of business processes according to regions (countries, regions, etc.); the need to promptly, regularly and periodically receive the necessary content; time-saving to obtain the desired content; personalization in service in the Web systems; with integration different ECCS [1-2]. Advantages of Web systems implementation are to increase the efficiency of content obtaining; reducing the cycle of production and sales; reducing costs associated with the information exchange; openness about users; automatically informing users about the content; creation of alternative sales channels such as the newspapers or online logs in Internet [1-2, 6, 10-15].

3 Recent research and publications analysis

Progress in the Web content leads production to a decrease in the general awareness level of the potential user. Increased content leads to the impossibility of his immediate processing and it's spread speed. In addition there was also a specific problems number (Table 1) [1-2, 6, 10-15].

Table 1. The Web content forming main problem

Name	Rationale	Rationale
Information noise	Content arrays structured.	Filters, content monitoring, site analysis, content analysis.
Parasitic content	Appearance as applications.	Filters, content monitoring, content analysis.
No content relevance	User needs inconsistency.	Create annotated database, primary content images search and his clustering, content analysis.
Content duplication	Content repeating in information sources.	Content analysis, scanners and filters based on statistics and criteria.
Navigation in a content stream	Rapid growth of the amount and content distribution.	Site analysis, filters, content monitoring, content analysis.
Search result redundancy	Duplication and no relevance.	Annotated search, content analysis and abstracting.

The reason for the loss of relevance of traditional information retrieval systems is scope/relevance growing fast and irregular dynamics of content streams (constant systematic or regular content updates). Large dynamic content streams coverage and summarize requires qualitatively new methods/approaches for problems solving of content creating and processing [7-10]. The content-monitoring software application provides the ability to automate finding the most important components in the Web content sources /streams. Their application caused needs by the systematic tracking of trends and processes in the Web content environment that is constantly updated. Web content monitoring is meaningful analysis of content streams. It is necessary to continuously obtain of the necessary qualitative and quantitative content sections within pre-undetermined time period [9, 12, 14]. Content monitoring components is the content search and content analysis [1, 4, 5, 7-11, 15, 16]. Content search is the operations set that required for finding in the predefined content sources. It is matches a user query in natural language [1-2, 30-35]. Working with text in natural language is challenging for mathematical linguistics. From solving the morphological analysis problem work of text implemented within sentences or text in natural language, as linearly ordered set of sentences, phrases or words.

Great importance is a linguistic unit presence/absence for automatic content-retrieval and textual content processing. Also, great importance is a particular category occurrence frequency of linguistic units in test Web content array [15, 16]. Quantitative calculation allows us to objectively conclusions about content orientation by the analysis units used number (key quotes) in the studied areas. For example, sometimes it is important to find the positive/negative feedback number on a certain product type [9, 15, 16]. Qualitative analysis allows us to objectively conclusions about a presence desired the linguistic unit in Web content array and context direction [9, 15, 16]. Content search is performed not by text content. Search performed with the brief characteristics of text that search content pattern (SCP). Here the main Web content is served in terms of specialized textual retrieval language [8-14]. SCP determination procedure is the main content text indexing, semantic analysis and translating it into text retrieval language (Table 2) [8, 12, 14]. The module does not retain

content text and it's SCP. For search of indexed Web content used content analysis to textual requests. An information request is the search order (SO), if it translated into information retrieval language and additional data complemented for finding [8, 12, 14]. Indexing depth is the content presentation detail degree in SCP for his central theme/subject, and related topics/subjects. Automating this process provides its unification and freeing some personnel from unproductive labour of Web content indexing [12-16]. Content-search contains semantic tools set: information retrieval language, content/queries indexing and search methods [12-16].

Table 2. The stages of content search

Operation	Operation description
SCP formation	SCP creation, administration, storage in modules.
Requests and SO formation	User requests and SO creation, administration and storage in module.
Search for content	CSI comparison of user request SO.
Content analysis	Quantitative and qualitative textual content analysis.
Result forming	The applying Web content analysis result is positive in the range (0.7, 1] or (0.5, 1].
Decision making	The decision on issuing Web content according to the applying Web content analysis result.
Content submission	Web content pretend that meets user information request.

Content formatting is the indexing, semantic analysis, main content determination of text and it convert into XML-format process. Formatting Web content performed the moderator manually or automatically by content analysis means [9]. In information retrieval language among the major elements (Table 3) does not use synonyms and homonyms through their semantic ambiguity [1-5, 11-16].

Table 3. Basic elements of information retrieval language

Element name	Language elements characteristic
Alphabet	The graphic characters set for language words and expressions fixing.
Lexis (paradigmatic)	A linguistic units related set, i.e. words used in the language.
Fragments blocks	This is a between phrase unity set that provide integrity through text content and thematic connections.
Grammar (syntagmatic)	The rules set for combining linguistic units in phrases, i.e. sentences constructing effective means.
Between phrase matching unity	The statements set, united semantically and syntactically in the sample. The core unity is the expression that is not subject to another statement and retains meaning in the context allocation.
Sentences	Implemented sentence, i.e. statement is a sentence in natural language, but the reverse is not true.
Paradigms	Words lexical-semantic group with a subject-logical relations based on semantic features.
Syntagmatic relation	Linear relation between words, which are set when words combining into phrases and phrases.
Paradigmatic	Relation basic and analytical between words that do not depend on the

Element name	Language elements characteristic
relation	context in which they are used, and connections caused not linguistic and logical.
Indexes identification rules	This is a language paradigmatic (vocabulary) and syntagmatic (grammar).

When information retrieval languages developing pay attention to the following aspects: specific sector/theme for which it is developed; texts features in the search content array; the information needs nature of users in Web systems [12-13].

The initial information in process operation of Web systems is evidence of appointment and conditions in the system. They define the main purpose simulation Web systems. They also make it possible to formulate the requirements for systems S and content processing subsystems. [1-2, 6, 10-15] Logistic Web systems model is $S = \langle X, C, V, H, Function, T, Y \rangle$, which $X = \{x_1, x_2 \ldots, x_{n_X}\}$ is entrance effects on the system; $Q = \{q_1, q_2 \ldots, q_{n_Q}\}$ is the users impact on the system; $C = \{c_1, c_2 \ldots, c_{n_C}\}$ is the content impact on the system; $V = \{v_1, v_2 \ldots, v_{n_V}\}$ is the external environment; $H = \{h_1, h_2 \ldots, h_{n_H}\}$ is the internal parameters of the system; $Z = \{z_1, z_2 \ldots, z_{n_Z}\}$ is information resources components of system; $T = \{t_1, t_2 \ldots, t_{n_T}\}$ is time transaction of the content processing; $Y = \{y_1, y_2 \ldots y_{n_Y}\}$ is output characteristics of the system [11]. The quantities x_i, c_r, v_l, h_k, y_j are the elements of disjoint subsets. They contain deterministic and stochastic components [11]. The process of Web systems S operation described by the function $y_j(t_i + \Delta t) = Function(x_i, c_r, v_l, h_k, t_i)$ [1-2, 11], where x_i is the requests for information of visitors/users to Web systems. According to Google Analytics from [6] y_j is the number of visits per time period Δt; the average time on site (min: c); rate of failures (%); achieved goal; dynamics (%); the number of all browsing; the number of page views for each visit; new visits (%); absolute unique visitors; traffic sources in % (search engines, direct traffic or other sites). Effects of values c_r, v_l, h_k on y_j to Web systems is unknown and unexplored [1-2, 6, 10-15].

4 Problems selection

Text content (book, articles, comments, etc.) contains a significant amount of a text in natural language, where the information is abstract. Text is united for the content lexical items sequence, the basic properties which are structural, informational, communicative coherence and integrity. They reflect the Web content and structural essence of text [1-2]. As functional-semantic structural unity of the text has construction rules, discovers patterns and formal connection meaningful of constituent units [1-3, 13]. Text connectivity is determined by external structural parameters and formal dependence of textual components, and integrity – through conceptual, modal and thematic dependence [3]. The outside (composition) and the internal textual structure is determined text units by internal organization patterns and their relationship to the text as a whole content. At the compositional level isolated sentences, paragraphs,

paragraphs, chapters, sections, subsections, pages, etc., but the sentence indirectly related to the internal structure, and therefore not considered [2-3, 13]. When using of statistical analysis methods ignore natural language the linguistic interconnectedness and non-linearity. The intermediate levels no involvement of text representation in the form structures semantic explain efficient formalism describing lack for the text structure. The semantic-grammatical (syntactic) and compositional level units of text are in the relationship and interdependence, in some cases identical, superimposed on each other (e.g. unity between phrase matching and paragraph though while they retain distinctive features). The semantic, grammatical and compositional structure of text associated with its style and stylistic characteristics. Each text reveals functional-stylistic orientation (scientific, artistic, etc.) and has stylistic qualities dependent on the author individuality and the text orientation [2-3, 13]. This complicates the content formatting process from different authors. The Web systems model does not reveal the mechanisms of content processing [2]. Formal logistic models of Web content management is assignment for determining the aging (relevance) process of content stream. Some of them (logistics, analytical) are also intended to the thematic flow analyze [2-4, 6-8]. They do not solve the problem of the Web content formation and support. They also solve the problem of not all content management, such as content submission plurality of end user according to his request, history or information portfolio; thematic content identification; automatic generation of digests and information portraits; content relationship tables building; content ratings calculation; information gathering from various sources and formatting; keywords/concepts identification; content duplicates finding; content categorization; selective dissemination of content [1-8, 10, 12]. Disadvantage of Web content management models is a connections lack between input data, Web content and output data in the Web systems [2].

5 Goals formulation

The purpose of the paper is the functional logistic model creation of Web content processing for the information resources formation in Web systems. The work relevance is the need to operational/objective assessment obtain of competition in the financial market segment of Web content and evaluate the financial market competitiveness of the content distribution. In the paper we will examine the stages of information resources processing and develop an optimal lifecycle for content processing. Implementing of functional logistic model of Web content processing enables a means of information resources formation in Web systems.

6 Research results analysis

The main subsystems of information resources processing in Web systems are the content formation, management and support, the circuit connections which is as follows [1-2, 6, 10-15]: *Web content formation* $\rightarrow$ *Web content management* $\rightarrow$ *Web content support.* Logistic model of Web systems presented as

$$S = \left\langle \begin{matrix} X, Q, Formation, H, C, V, \\ Management, Support, Z, T, Y \end{matrix} \right\rangle \quad (1)$$

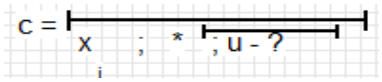
where the value $X = \{x_1, x_2 \ldots, x_{n_X}\}$ is set of input data $x_i \in X$ from different sources at $i = \overline{1, n_X}$; the value $Q = \{q_1, q_2 \ldots, q_{n_Q}\}$ is set of user queries $q_d \in Q$ while $d = \overline{1, n_Q}$; the value *Formation* is the operator of content formation; the value $H = \{h_1, h_2 \ldots, h_{n_Y}\}$ is set of internal parameters $h_k \in H$ of the system S when $k = \overline{1, n_H}$; the value $C = \{c_1, c_2 \ldots, c_{n_C}\}$ is set of Web content $c_r \in C$ at $r = \overline{1, n_C}$; the value $V = \{v_1, v_2 \ldots, v_{n_V}\}$ is set of the influence parameters $v_l \in V$ of the environment on the system S at $l = \overline{1, n_V}$; the value *Management* is the operator of content management; the value *Support* is operator of Web content support; value $Z = \{z_1, z_2 \ldots, z_{n_Z}\}$ is set of information resource pages $z_w \in Z$ of in the system S at $w = \overline{1, n_Z}$; the value $T = \{t_1, t_2 \ldots, t_{n_T}\}$ is time $t_p \in T$ transaction of information resource processing in the system S when $p = \overline{1, n_T}$; the value $Y = \{y_1, y_2 \ldots y_{n_Y}\}$ is set of statistical data $y_j \in Y$ in system S at $j = \overline{1, n_Y}$. Web systems operation described such relationships schemes of its main components [1-2, 6, 10-15]:

- to information resource formation of system scheme is such *data processing* → *Web content formation* → *Web content database* → *Web content management* → *information resource processing in Web systems*;
- for the answer formation to a user's query schema is such *the user query processing* → *Web content management* → *information resource processing in Web systems* → *Web content support* → *database users processing*;
- to a report create of the system operation to moderator scheme is such *moderator requested* → *Web content support* → *user database processing* → *Web content management* → *report formation*;
- for the internal parameters moderating of the system scheme is such *requested* → *Web content formation* → *base rules formation or processing* → *Web content support* → *base rules formation or processing* → *Web content management* → *result formation*.

The content formation described by the operator form $c_r = Formation(u_f, x_i, t_p)$, where u_f is conditions set for the content formation, i.e. $u_f = \{u_1(x_i), \ldots, u_{n_U}(x_i)\}$. Content is presented as follows:

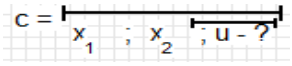
$$c_r = \left\{ \bigcup_f u_f \left| \begin{matrix} (x_i \in X) \wedge (\exists u_f \in U), \\ U = U_{x_i} \vee U_{\overline{x_i}}, i = \overline{1, m} \end{matrix}, f = \overline{1, n} \right. \right\}, \quad (2)$$

so, the operation performed elimination of term x_i and empty term $*$ selection accordance to execute the specified conditions u [9]:



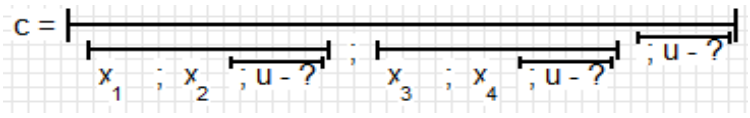
(3)

Then elimination operation [9] of term choice between x_1 and x_2 presented as



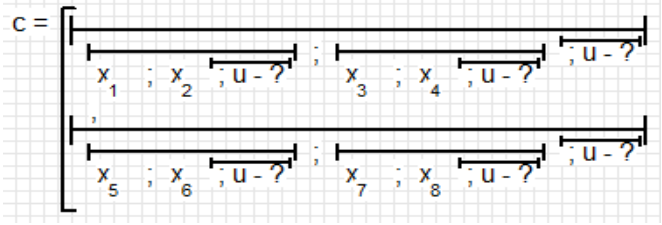
(4)

The elimination operation [9] with the increase of the input data x_i from different sources is as equation (5).



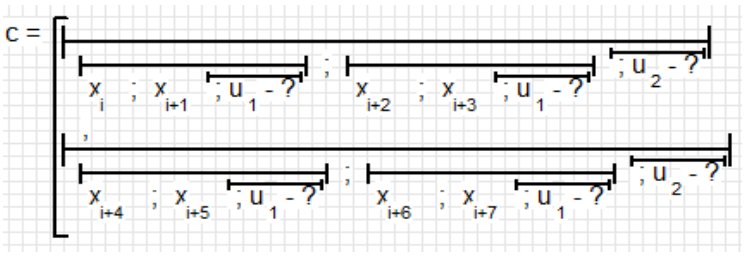
(5)

Content c_r is formed by elimination in (6) from different data sources according to the specified performance conditions u [1-2, 6, 10-15]. That is apply filter rules in (6) to select of relevant content from multiple of found content from relevant sources or created by moderators.

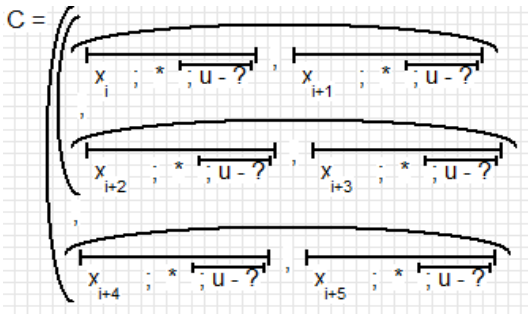


(6)

In this case there is the possibility of empty content c_r. It is therefore necessary to introduce refined set of elimination conditions u_f [9]. It is also necessary perform this operation in parallel with (7) for different input data. Sequencing operation (8) regulates the content c_r formation.

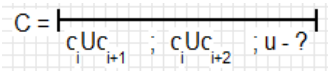


(7)

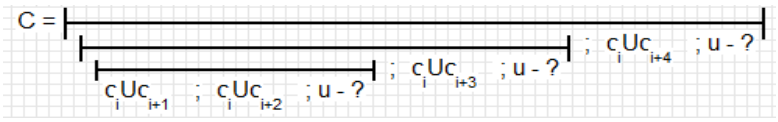


(8)

Elimination operation (9) contributes to a set formation of content C. Aggregate apply of elimination operations to multiple select of content C promotes a set formation of relevant content. That is applying filter rules to the content set and not relevant content clipping in (10).

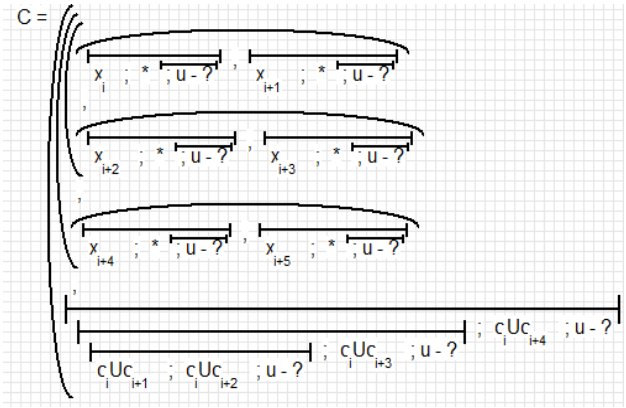


(9)



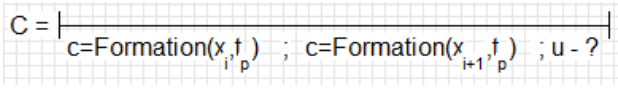
(10)

Equation (11) reveals the multiple collecting process of content C from different sources of data. Here take into account the filter rules to identify not relevant content and content that does not comply with the requirements for the Web content formation in the Web systems.



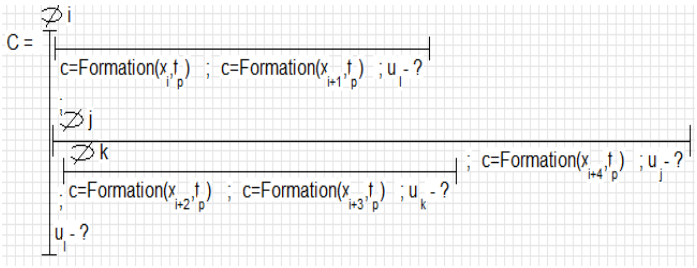
(11)

Elimination operation forming a plurality of Web content is (12). Repeat elimination in (12) can start the set forming process of relevant content recursively depending on the time setting t_j [1-2, 6, 10-15]. This option is set by the moderator in Web systems of set $T = \{t_1, t_2 \ldots, t_{n_T}\}$ for each source of data separately.

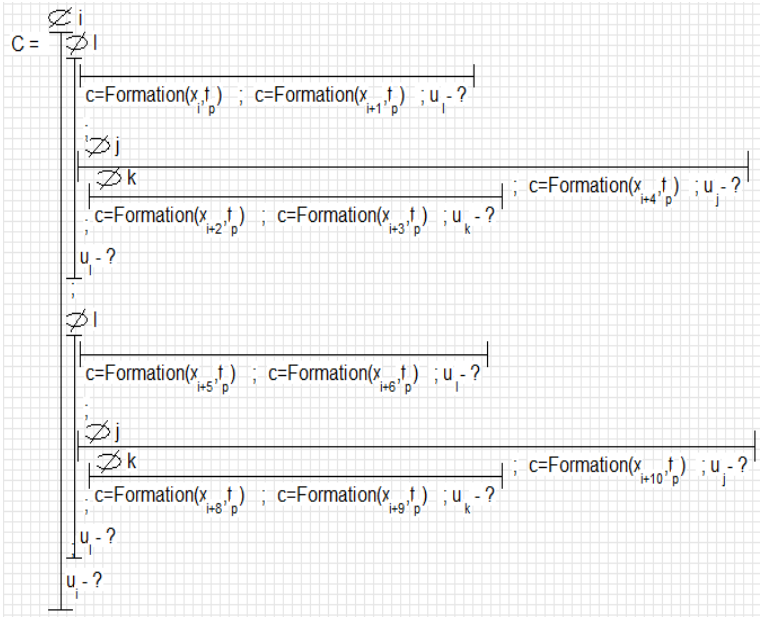


(12)

Sub cyclic elimination in (13) generates a relevant Web content set without duplication from a limited list of information sources. The relevant content multiple formation is recursive, depending on the conditions u_i of data formation and filtering. These conditions are specified by moderators in Web systems of rules set $U^i = \{u_1, u_2 \ldots, u_{n_U}\}$ for each source of data separately and common rules set $U = \{u_1, u_2 \ldots, u_{n_U}\}$ for all sources by previous content analyzing from these sources. Cycle sequencing application to cyclic elimination operations in (14) for the Web content formation can extend the list of information sources and more accurately stage of content duplication identifying. The relevant content multiple formation is recursive, depending on the time setting t_p and the conditions u_i of data formation and filtering.



(13)

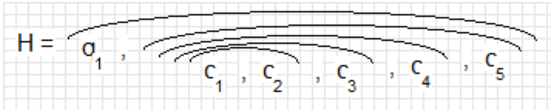


(14)

Web content management step describes the operator form $z_w = Management(q_d, c_r, h_k, t_p)$, where $Q = \{q_1, q_2 \ldots, q_{n_Q}\}$ is the set of user queries; h_k is a set of conditions of Web content management, i.e. $H = \{h_1(c_{i+1}, q_d), \ldots, h_{n_H}(c_{i+n_H}, q_d)\}$. Management of Web content presented as

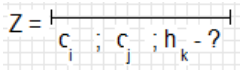
$$z_w = \left\{ \bigcup_{k=1}^{n_H} h_k(c_{i+1}, q_d) \middle| \begin{array}{l} (c_{i+k} \in C) \wedge (q_d \in Q) \wedge (h_k \in H_q), \\ H = H_{q_d} \vee H_{\overline{q_d}}, i = \overline{1, n_C}, \\ d = \overline{1, n_Q}, k = \overline{1, n_H} \end{array} \right\},$$

so the use of sequencing operations to relevant content select c_r from content multiple $C = \{c_1, c_2 \ldots, c_{n_C}\}$ only by user request $q_d \in Q$:



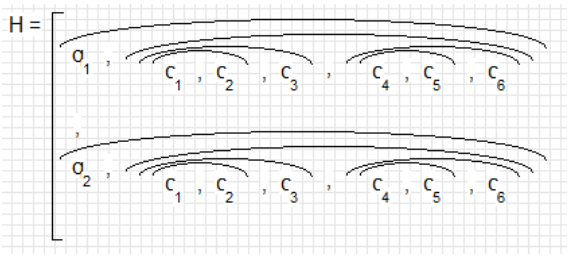
(15)

Elimination operations application to content select appropriate conditions $v_l = (v_1(q_i, h_k, c_r, z_w, t_p), \ldots, v_{n_V}(q_i, h_k, c_r, z_w, t_p))$ is affect the process of information resources formation by user request:



(16)

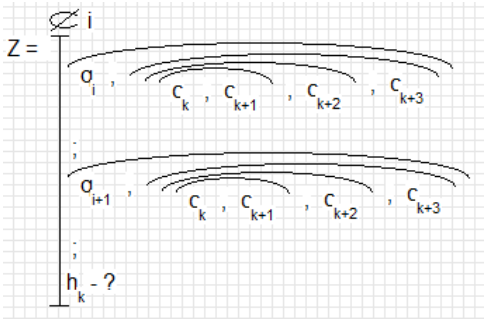
Set formation $H = \{h_1, h_2 \ldots, h_{n_Y}\}$ of content management conditions describes the parallelization operation of Web content analysis c_r according to user request $q_d \in Q$ in Web systems:



(17)

Cycle sequencing (18) generates a pages plurality $Z = \{z_1, z_2 \ldots, z_{n_Z}\}$ of information resource in accordance with the content analysis of c_r Web content sets, q_i user queries in Web systems and content management conditions h_k. Cyclic sequencing and elimination application in (17) of similar queries for users and for pages multiple caching $Z = \{z_1, z_2 \ldots, z_{n_Z}\}$ forms the resource according to the content analy-

sis of Web content c_r sets, users' queries q_i in Web systems and conditions h_k of Web content management.



(18)

The content accompany step *Support* described in the operator form

$$y(t_p + \Delta t) = Support(v_l, h_k, c_r, z_w, t_p, \Delta t), \tag{19}$$

where v_l is conditions set of content support and external influences of environment on the system, i.e. $v_l = (v_1(q_i, h_k, c_r, z_w, t_p), \ldots, v_{n_V}(q_i, h_k, c_r, z_w, t_p))$. Output statistics implemented are as

$$y_j = \left\{ \bigcup_l v_l \middle| \begin{array}{l} (\exists q_d \in Q) \wedge (\exists z_w \in Z) \wedge \\ \wedge (\forall v_l \in V) \wedge (\forall (c_r \wedge q_d) \in h_k), \\ V = V_{q_d} \vee V_{\overline{q_d}}, d = \overline{1, n_Q}, l = \overline{1, n_V}, \\ w = \overline{1, n_Z}, r = \overline{1, n_C}, k = \overline{1, n_H} \end{array} \right\} \tag{20}$$

The following equation reflects the Web content support on the Web systems. That is describes the process of the users reactions analyzing for information resources processing in such systems. Functional logistic model describes the Web systems operation process with basic processes of information resources processing as the content formation, management and maintenance. Statistics analysis of the Web systems operation is conducted by regular/potential user reactions analysis to the system (visit, requests, search for keywords, etc). Application of elimination, sequencing and parallelization operations in the proper order facilitates effective analysis of target/potential audience's response for the Web systems functioning. It also helps to predict relevant changes in demand for Web content. The general design principles of Web systems architecture allow implementing process of information resources processing to expand the similar systems functionality.

7 Web content formation model and method

The Web content formation for Web site provides a link between the input data from

different Web sources set and the Web content set into the appropriate database in Web systems that can be presented as $Sourse(x_i) \rightarrow x_i \rightarrow X \rightarrow Formation(u_f, x_i, t_p) \rightarrow c_r \rightarrow C \rightarrow DataBase(C)$, where $Sourse(x_i)$ is Web content source, x_i is matched content from the source, X is the relevant sources data set, $Formation(u_f, x_i, t_p)$ is Web content formation operator in a fixed time t_p under u_f appropriate conditions, c_r is formed Web content under u_f conditions, C is generated Web content set, $DataBase(C)$ is Web content prevailing database. Web content formation model in Web systems can be showed as

$$Formation = \left\langle \begin{array}{l} X, Gathering, Formatting, \\ KeyWords, Backup, \\ Caterization, BuDigest, \\ Dissemination, T, C \end{array} \right\rangle,$$

where $X = \{x_1, x_2 \ldots, x_{n_X}\}$ is input data set $x_i \in X$ from different Web resources or the moderators at $i = \overline{1, n_X}$; $KeyWords$ is Web content keywords and concepts identify operator; $Formatting$ is Web content formatting operator; $Gathering$ is Web content collecting/creating operator from various Web sources; $Categorization$ is Web content categorization operator; $Backup$ is Web content duplicate detect operator; $BuDigest$ is Web content digest formation operator; $Dissemination$ is Web content selective distribution operator; $T = \{t_1, t_2 \ldots, t_{n_T}\}$ is Web content forming transaction time $t_p \in T$ while $p = \overline{1, n_T}$; $C = \{c_1, c_2 \ldots, c_{n_C}\}$ is Web content set $c_r \in C$ with $r = \overline{1, n_C}$. The Web content formation is described by the form $c_r = Formation(u_f, x_i, t_p)$ operator, where u_f is Web content formation conditions set, i.e. $u_f = \{u_1(x_i), \ldots, u_{n_U}(x_i)\}$. Web content submitted as follows:

$$c_r = \left\{ \bigcup_f u_f \left| \begin{array}{l} (x_i \in X) \wedge (\exists u_f \in U), \\ U = U_{x_i} \vee U_{\overline{x_i}}, i = \overline{1, m} \end{array}, f = \overline{1, n} \right. \right\},$$

that the data set convert following steps passing in a relevant, formatted, classified and validated content set: $x_i \in X \rightarrow Gathering(u_f, x_i, t_p) \rightarrow Backup(c_r, u_b, t_p) \rightarrow Formatting(c_r, t_p) \rightarrow KeyWords(c_r, t_p) \rightarrow Categorization(c_r, t_p) \rightarrow BuDigest(c_r, t_p) \rightarrow Dissemination(c_r, t_p) \rightarrow c_r \in C$.

Decisions that can help to navigate in the dynamic input data from different Web sources, provide the data syndication $C = Gathering(X, U_G, T)$, where $Gathering$ is Web content collecting/creating operator, X is Web content set from data different Web sources, U_G is data collecting conditions set from various Web sources, T is the content collection/creation time. Content duplicate detecting marked by the operator as $C = Backup(Gathering(X, U_G, T), U_B)$, where $Backup$ is the text content duplication identify operator, X is Web content set from data different Web sources,

U_B is text content duplication identify conditions set, C is Web content set. Content duplicate identifying in text is based on the linguistic statistical methods for general terms detecting, which a content form the verbal signature chain.

Content syndication technology contains data collect programs learning process with the individual sources structural characteristics (from journalists, visitors, editors, users, moderators, different Web resources,), Web content direct scanning and bringing the total: $C = Formatting(Backup(Gathering(X, U_G, T), U_B), U_{FR})$, where $Gathering$ is the content collecting/creating operator, $Formatting$ is Web content formatting operator, U_G is data collecting conditions set from various sources, U_{FR} is information formatting conditions set, T is the content collection time.

Keywords identify defined by the operator $KeyWords(C)$ as: $C = KeyWords(Formatting(Backup(Gathering(X, U_G, T), U_B), U_{FR}), U_K)$, where $KeyWords$ is the content keywords and concepts identify operator that is implemented as a processes set, using the presented chart in Fig. 1; $Formatting$ is content formatted operator; U_G is data collecting conditions set from various sources; $Gathering$ is the content collecting/creating operator; U_{FR} is conditions data formatting set; T is the content collection time; U_K is keywords and concepts identify conditions set.

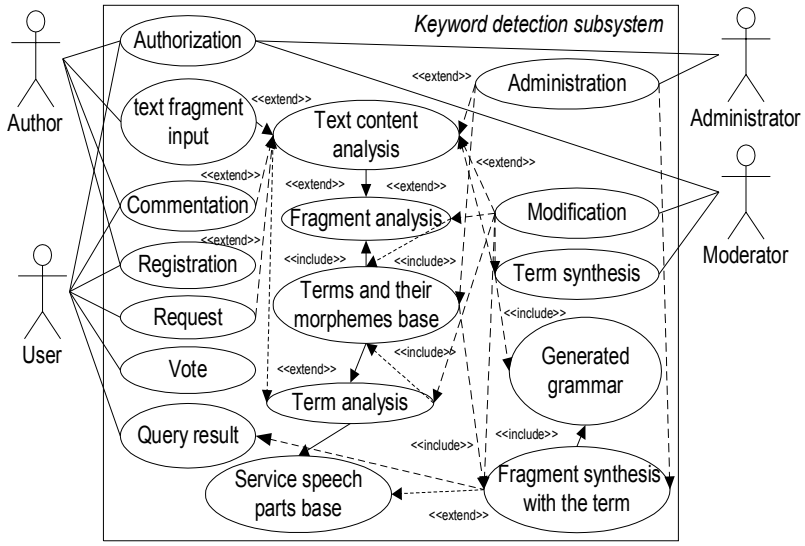


Fig. 1. Use case diagram for the content keywords identifying process

Terms searching is defined using terms/morphemes database, speech service part database and text analysis rules. Based on the generated grammar rules perform correction term according to its use in context. Classification and content distribution means is an information retrieval system for content selective distribution (Content Router). Content analysis for compliance thematic requests to

$C_{Ct} = Categorization\left(KeyWords(C,U_K),U_{Ct}\right)$, where U_{Ct} is categorization conditions set, C_{Ct} is rubrics relevant content set. Digest set C_D formed by such dependence as $C_D = BuDigest\left(C_{Ct},U_D\right)$, where U_D is conditions set for the digests formation, $BuDigest$ is digests forming operator, i.e.

$$C_D = BuDigest\left(Categorization\left(KeyWords(C,U_K),U_{Ct}\right),U_D\right).$$

Web content selective distribution described as $C_{Ds} = Dissemination\left(C_D,U_{Ds}\right)$, where C_{Ds} is content selectively distributed set, U_{Ds} is content selective distribution conditions set, $Dissemination$ is the content selective distribution operator. In Fig. 2, a submitted cooperation diagram for content subject keywords identifying process.

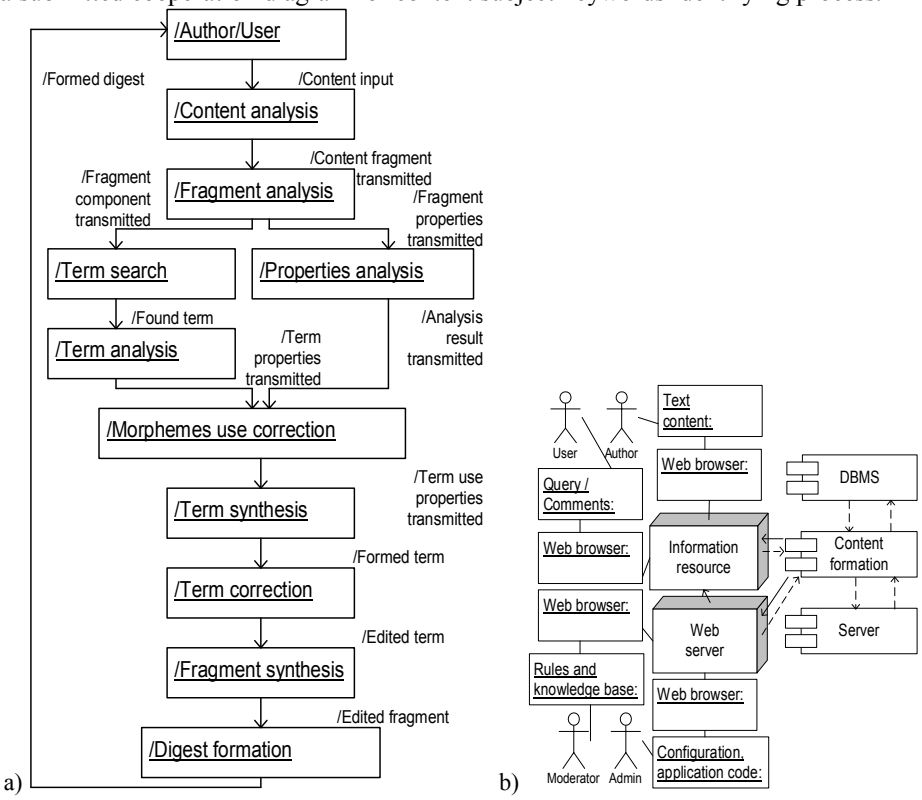


Fig. 2. a) Cooperation diagram and b) components diagram for the process to content subject keywords find

The text lexical-grammatical and semantic-pragmatic construction analysis used in Web content automatic categorization, whose main task is to find text in the content flow through the content analysis that best matches the content topics and user needs. After text fragment and term analyzing is the new term synthesis as a content topic keyword. In Fig. 2, b submitted component diagram for content topic keyword process. In Fig. 3 activity diagram for content subject keywords identifying process is showed. The present method next step in content forming is content categorization.

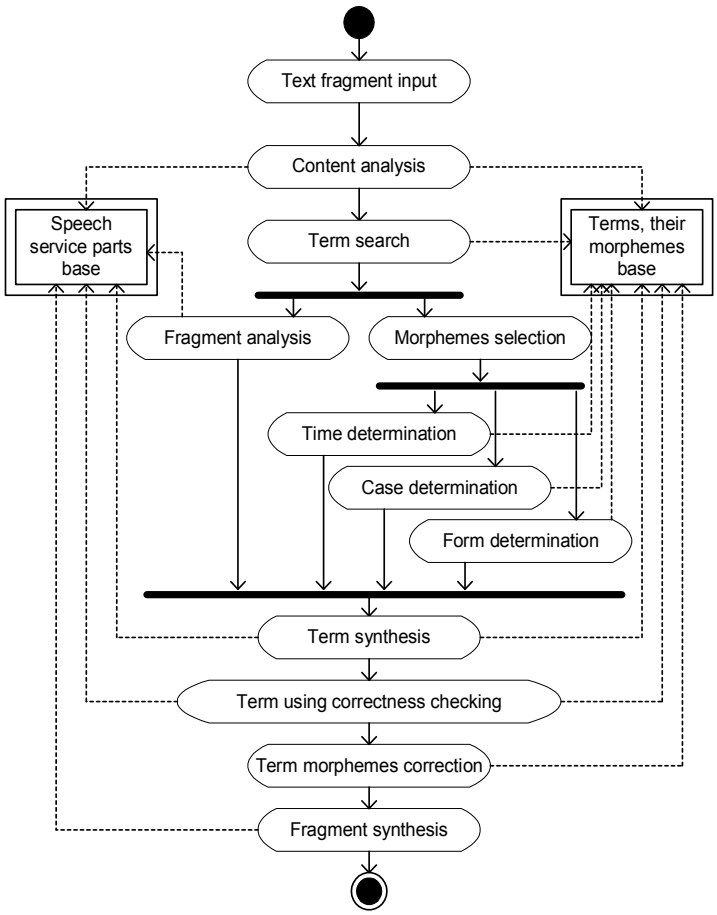


Fig. 3. Activity diagram for the content subject keywords identifying process

Based on the developed method content forming subsystems at various stages are implemented in Internet projects vgolos.com.ua (1), autochip.vn.ua (2), press-time.com.ua (3), fotoghalereja-vysocjkykh.com (4), kursyvalyut.com (5), tatjana.in.ua (6), dobryjranok.com (7). Table 4 presents the developed systems comparative characteristics derived from Google Analytics.

Table 4. The system work comparative characteristics for the time period from 10.2012 till 11.2017 years

Systems characterization	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Content formation	+	+	+/-	+/-	+	-	+/-
New Visits (%)	51,83	77,59	45,65	80,82	90,87	36,43	12,07
Returning Visitor (%)	48,15	77,59	54,35	82,19	62,79	63,27	87,93
Unique visitors	217 719	326	123 756	62	42	21	7
Visiting	326 940	406	167 856	73	103	49	58
Pages/Visit	1,72	2,13	1,45	1,86	1,67	2,90	3,90

Systems characterization	1	2	3	4	5	6	7
Pages browse	562 455	863	345 234	136	237	142	226
Fault indicator (%)	76,92	56,90	79,56	71,23	61,23	46,94	48,28
Visits average duration	01:45	01:08	01:09	00:47	00:37	04:38	09:35
New Visitor (%)	51,85	22,41	45,65	17,81	37,21	36,43	12,07
Visitors from Ukraine (%)	89,81	73,89	92,33	87,67	97,07	71,43	55,17
Visitors from Russia (%)	2,55	17,00	6,27	2,74	1,05	24,49	43,10
Visitors from the USA (%)	0,58	0,05	0,06	1,37	0,61	0,07	1,72
Traffic Conversion (%)	54,62	3,45	34,65	12,33	35,65	0	6,90
Search traffic (%)	36,03	88,67	60,05	69,86	59,03	73,47	43,10
Direct traffic (%)	9,21	7,88	5,25	17,82	5,32	26,53	50,00
Traffic campaigns (%)	0,14	0	0,05	0	0	0	0

Google Analytics provides advanced data analysis and allows us to estimate Web content traffic and marketing activities effectiveness, such as vgolos.com.ua (Fig. 4).

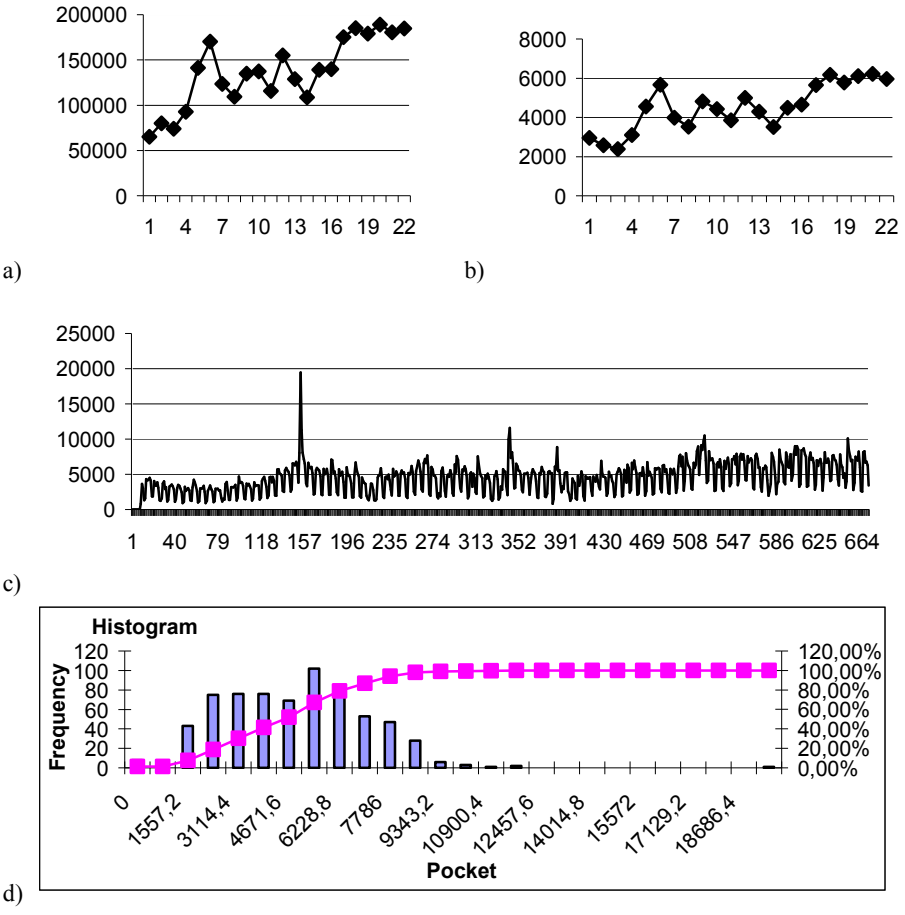


Fig. 4. Visitors distribution of a) total, b) medium and c) daily, d) monthly number in 2010-2017 years

The Web content formation model implement in the form of content-monitoring complexes to content collection from data various sources and provide a content database according to the users information needs. As a result, content harvesting and primary processing its lead to a single format, classified according to the Categories and he is credited tags with keywords. This facilitates the Web content management process implementation. In text analyzing explore its layered structure: the source text as a characters linear sequence; morphological structure linear sequence, statements linear sequence, related unity net. The text preliminary study provides for the text division into individual tokens that carry out the finite automata method. Entry information is text in natural language text as a characters sequence, and output information – analyzed text partition, sentences and tokens table. There is the following relationship: the more unique content in the Web systems, the more the visitor's information resource in its system. Web content formation subsystem reduces the time to fill out unique content an information resource and increases the volume in a short time at this unique content in information resources and the queries number from search engines. These data take into account when creating or updating information resource and improve the Web systems architecture.

8 Conclusion

In the given paper is functional logistic model of Web content processing in e-business systems developed. The model is based on the layered structure of processes. This model involves the division of the overall process into the following stages: content collection/creation from different sources; content formatting; keywords and concepts identifying; content categorization; content duplicate detection; digests formation; selective distribution of content between moderators and users of Web systems. The model is based on the principles of content analysis. It automates the various steps of information product creating of this type without loss of content and lower quality. The method effectiveness confirms the results of its application in developing a number of Web content projects. Developed automation Web content processing allows to speed up the content formation, management and maintenance process. It also contributes to the rating increase of generated by their use with Web information resources. The given article describes a Web content forming method based on processes multilevel models. This model involves the overall process division into the following stages: content collection/creation from different sources, formation, keywords and concepts identifying, categorization, duplicate identify, digests formation and content selective distribution between moderators and users in the Web systems. It is based on the content analysis principles, which allows you to automate various phases of this type information product creating without content loss and quality lower. The method effectiveness is its application the results confirming in developing a Web content projects number. Developed automation formation content allow you to speed up the content forming process and increase the use of ratings generated by them through Web resources.

References

1. Vysotska, V., Hasko, R., Kuchkovskiy, V.: Process analysis in electronic content commerce system. In: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT 2015, 120-123 (2015)
2. Vysotska, V.: Linguistic Analysis of Textual Commercial Content for Information Resources Processing. In: Modern Problems of Radio Engineering, Telecommunications and Computer Science, TCSET'2016, 709–713 (2016)
3. Gozhyj, A.: Development of Fuzzy Situational Networks with Time Constraints for Modeling Dynamic Systems, Naukovi visti KPI, vol. 5, 15-22 (2015)
4. Kravets, P.: The control agent with fuzzy logic, Perspective Technologies and Methods in MEMS Design, MEMSTECH'2010, 40-41 (2010)
5. Kravets, P., Kyrkalo, R.: Fuzzy logic controller for embedded systems. In: Proc. of the 5th Int. Conf. on Perspective Technologies and Methods in MEMS Design, (2009)
6. B. Clifton, Google Analytics: professional attendance analysis web sites, Williams, 2009.
7. Vysotska, V., Rishnyak, I., Chyrun L.: Analysis and evaluation of risks in electronic commerce, CAD Systems in Microelectronics, 9th International Conference, 332-333, (2007).
8. Mykich, K., Burov, Y.: Research of uncertainties in situational awareness systems and methods of their processing, EasternEuropean Journal of Enterprise Technologies, vol. 1(79), 19-26 (2016)
9. Mykich, K., Burov, Y.: Uncertainty in situational awareness systems, Modern Problems of Radio Engineering, Telecommunications and Computer Science, TCSET, 729-732 (2016)
10. Kanishcheva, O., Vysotska, V., Chyrun, L., Gozhyj, A.: Method of Integration and Content Management of the Information Resources Network. In: Advances in Intelligent Systems and Computing, 689, Springer, 204-216 (2018)
11. Naum, O., Chyrun, L., Kanishcheva, O., Vysotska, V.: Intellectual System Design for Content Formation. In: Computer Science and Information Technologies, Proc. of the Int. Conf. CSIT, 131-138 (2017)
12. Korobchinsky, M., Vysotska, V., Chyrun, L., Chyrun, L.: Peculiarities of Content Forming and Analysis in Internet Newspaper Covering Music News, In: Computer Science and Information Technologies, Proc. of the Int. Conf. CSIT, 52-57 (2017)
13. Lytvyn, V., Vysotska, V., Burov, Y., Veres, O., Rishnyak, I.: The Contextual Search Method Based on Domain Thesaurus. In: Advances in Intelligent Systems and Computing, 689, 310-319 (2018)
14. Vysotska, V., Chyrun, L., Chyrun, L.: The Commercial Content Digest Formation and Distributional Process. In: Computer Science and Information Technologies, Proc. of the XI–th Int. Conf. CSIT'2016, 186-189 (2016)
15. Vysotska, V., Chyrun, L.: Analysis features of information resources processing. In: Computer Science and Information Technologies, Proc. of the Int. Conf. CSIT, 124-128 (2015)
16. Basyuk, T.: The main reasons of attendance falling of internet resource. In: Computer Science and Information Technologies, Proc. of the X-th Int. Conf. CSIT'2015, 91-93 (2015)
17. Basyuk, T.: The Popularization Problem of Websites and Analysis of Competitors. Advances in Intelligent Systems and Computing II. CSIT 2017. Advances in Intelligent Systems and Computing, vol 689. Springer, Cham pp. 54-65 (2017)
18. Davydov, M., Lozynska, O.: Information System for Translation into Ukrainian Sign Language on Mobile Devices. In: Computer Science and Information Technologies, Proc. of the Int. Conf. CSIT, 48-51 (2017).

19. Davydov, M., Lozynska, O.: Linguistic Models of Assistive Computer Technologies for Cognition and Communication. In: Computer Science and Information Technologies, Proc. of the Int. Conf. CSIT, 171-175 (2017)
20. Khomytska, I., Teslyuk, V.: Specifics of Phonostatistical Structure of the Scientific Style in English Style System. In: Computer Science and Information Technologies, Proc. of the XI–th Int. Conf. CSIT'2016, 129-131. (2016)
21. Khomytska, I., Teslyuk, V.: The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspaper Styles on the Phonological Level. In: Advances in Intelligent Systems and Computing, 512, 149-163 (2017)
22. Kowalik, Dagmara: Polish vocational competence standards for the needs of audult education and the Europen labour marke. In: International Conference on Advanced Information Engineering and Education Science, ICAEES, 95-98 (2013)
23. Rusyn, B., Lutsyk, O., Lysak, O., Lukeniuk, A., Pohreliuk, L.: Lossless Image Compression in the Remote Sensing Applications In: Proc. of the IEEE First Int. Conf. on Data Stream Mining & Processing (DSMP), 195-198 (2016)
24. Chernukha, O., Bilushchak, Y.: Mathematical modeling of random concentration field and its second moments in a semispace with erlangian disrtibution of layered inclusions. In: Task Quarterly, 20(3), 295-334 (2016)
25. Maksymiv, O., Rak, T., Peleshko, D.: Video-based Flame Detection using LBP-based Descriptor: Influences of Classifiers Variety on Detection Efficiency. In: International Journal of Intelligent Systems and Applications, 9(2), 42-48 (2017)
26. Teslyuk, V., Beregovskyi, V., Denysyuk, P., Teslyuk, T., Lozynskyi, A.: Development and Implementation of the Technical Accident Prevention Subsystem for the Smart Home System. In: International Journal of Intelligent Systems and Applications, 10(1), 1-8 (2018)
27. Tkachenko, R., Tkachenko, P., Izonin, I., Tsymbal, Y.: Learning-based image scaling using neural-like structure of geometric transformation paradigm. In: Studies in Computational Intelligence, 730, Springer Verlag, 537–565 (2018)
28. Peleshko, D., Rak, T., Izonin, I.: Image Superresolution via Divergence Matrix and Automatic Detection of Crossover. In: International Journal of Intelligent Systems and Application, 8(12), 1-8 (2016)
29. Rashkevych, Y., Peleshko, D., Vynokurova, O., Izonin, I., Lotoshynska, N.: Single-frame image super-resolution based on singular square matrix operator. In: IEEE 1th Ukraine Conference on Electrical and Computer Engineering (UKRCON), 944-948 (2017)
30. Su, J., Vysotska, V., Sachenko, A., Lytvyn, V., Burov, Y.: Information resources processing using linguistic analysis of textual content. In: Intelligent Data Acquisition and Advanced Computing Systems Technology and Applications, Romania, 573-578, (2017)
31. Chen, J., Dosyn, D., Lytvyn, V., Sachenko, A.: Smart Data Integration by Goal Driven Ontology Learning. In: Advances in Big Data. Advances in Intelligent Systems and Computing. – Springer International Publishing AG 2017. P. 283-292 (2017).
32. Kravets, P.: The game method for orthonormal systems construction. In: The Experience of Designing and Application of CAD Systems in Microelectronics (2007).
33. Lytvyn, V., Vysotska, V., Peleshchak, I., Rishnyak, I., Peleshchak, R.: Time Dependence of the Output Signal Morphology for Nonlinear Oscillator Neuron Based on Van der Pol Model. In: International Journal of Intelligent Systems and Applications, 10, 8-17 (2018)
34. Riznyk, V.: Multi-modular Optimum Coding Systems Based on Remarkable Geometric Properties of Space. In: Advances in Intelligent Systems and Computing,, 129-148 (2017)
35. Lytvyn, V., Vysotska, V., Uhryn, D., Hrendus, M., Naum, O.: Analysis of statistical methods for stable combinations determination of keywords identification. In: Eastern-European Journal of Enterprise Technologies, 2/2(92), 23-37 (2018)