*commercial content, information resource, content analysis, content search, content monitoring, content lifecycle, electronic content commerce system*

*Victoria VYSOTSKA* [*], *Lyubomyr CHYRUN* [**]

# SET-THEORETIC MODELS AND UNIFIED METHODS OF INFORMATION RESOURCES PROCESSING IN E-BUSINESS SYSTEMS

**Abstract**

*In the given article is functional logistic model of commercial content processing as the content life cycle stage in electronic commerce systems proposed. The model of commercial content processing describes the information resources forming in electronic content commerce systems and automation technology simplifies the commercial content management. In the given article the main problems of electronic content commerce and functional services of commercial content processing are analyzed. The proposed model gives an opportunity to create an instrument of information resources processing in electronic commerce systems and to implement the subsystem of commercial content formation, management and support.*

## 1. INFORMATION

The Internet active development promotes the needs growth in production/strategic data and new forms of information services implementation [1-5, 7-9]. Documented information is an informational product or commercial content, if it is prepared in accordance with user needs and intended to meet them [2, 13-23]. Electronic content commerce systems development and implementation is one of the e-business development strategic directions. A characteristic feature of such systems is the automatic information resources processing to increase content sales of permanent user, for potential users active involvement and expanding the target audience boundaries [5]. Specialists in designing, implementation and deployment of electronic content commerce

[*] Information Systems and Networks Department, Lviv Polytechnic National University, S. Bandery Str., 12, Lviv, 79013, UKRAINE, +38 (032) 258-26-38, victana@bk.ru
[**] Software Department, Lviv Polytechnic National University, S. Bandery Str., 12, Lviv, 79013, UKRAINE, +38 (032) 258-25-78, chyrunlv@mail.ru

systems (ECCS) deal 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 [1, 2, 5, 9]. The special feature of ECCS is as follows [1-12]: 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 [2]. The modern market is characterized by an increase in demand for commercial content and growth proposals in public enterprises and public institutions [1-2]. Using commercial content helps optimize the management, trade and economic strategy and long-term development programs for production. It is associated with an increase in the complexity of management and implementation of systems using the predictive content of character. Specialized information resources are most type's users of commercial content, e.g., online newspapers, online magazines, online publishing, etc.

## 2. METHODOLOGY OF COMMERCIAL CONTENT PROCESSING

The actual problem in the electronic content commerce systems design, development, implementation and maintenance is to the research active development in the e-business. An important problem is the lack of theoretical justification, standardized methods and software for information resources processing in such systems. There are new approaches and solutions to this problem. But the important issue is the discrepancy between the known methods and software of information resources processing and the electronic content commerce systems construction principles. There is no common approach of electronic content commerce systems creation and standardized methods of information resources processing in these systems. One of the modernity main features is the constant growth rate of content production. This process is objective and positive. But there is one big problem. Progress in the 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 its spread speed. In addition there was also a specific problems number [2, 7-9, 12-23]. The initial information in ECCS operation process is evidence of appointment and conditions of the system. They define the main purpose simulation ECCS. They also make it possible to formulate the requirements for systems $S$ and content processing subsystems [2]. ECCS model is

$S = \langle X, C, V, H, Function, T, Y \rangle$, which $X$ – entrance effects on the system; $C$ – the content impact on the system; $Q$ – the users impact on the system; $V$ – the external environment; $H$ – the internal parameters of the system; $Z$ – information resources components of system; $T$ – time transaction of the content processing; $Y$ – output characteristics of the system [10]. The quantities $x_i$, $c_r$, $v_l$, $h_k$, $y_j$ are the elements of disjoint subsets. They contain deterministic and stochastic components [10]. The process of ECCS $S$ operation described by the function $y_j\ t_i + \Delta t = Function\ x_i, c_r, v_l, h_k, t_i$ [2, 10], where $x_i$ – the requests for information of visitors/users to ECCS. According to Google Analytics from [5] $y_j$ – the number of visits per time period $\Delta t$; 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 %. Effects for $c_r$, $v_l$, $h_k$ values on $y_j$ to ECCS is unknown and unexplored [2].

## 3. SET-THEORETIC MODELS OF INFORMATION RESOURCES PROCESSING IN E-BUSINESS SYSTEMS

The main subsystems of information resources processing in ECCS are the content formation, management and support, the circuit connections which is as follows [2]: *content formation* $\rightarrow$ *content management* $\rightarrow$ *content support.* Model of electronic content commerce systems presented as:

$$S = \langle X, Q, Formation, H, C, V, Management, Support, Z, T, Y \rangle, \qquad (1)$$

where: $X = \{x_1, x_2 \ldots, x_{n_X}\}$ – set of input data $x_i \in X$ from different sources at $i = \overline{1, n_X}$,

$Q = \{q_1, q_2 \ldots, q_{n_Q}\}$ – set of user queries $q_d \in Q$ while $d = \overline{1, n_Q}$,

*Formation* – the operator of content formation,

$H = \{h_1, h_2 \ldots, h_{n_Y}\}$ – set of internal parameters $h_k \in H$ of the system $S$ when $k = \overline{1, n_H}$,

$C = \{c_1, c_2 \ldots, c_{n_C}\}$ – set of commercial content $c_r \in C$ at $r = \overline{1, n_C}$,

$V = \{v_1, v_2 \ldots, v_{n_V}\}$ – set of the influence parameters $v_l \in V$ of the environment on the system $S$ at $l = \overline{1, n_V}$,

*Management* – the operator of content management,

*Support* – operator of commercial content support,
$Z = \{z_1, z_2 \ldots, z_{n_Z}\}$ – set of information resource pages $z_w \in Z$ of in the system $S$ at $w = \overline{1, n_Z}$ ,
$T = \{t_1, t_2 \ldots, t_{n_T}\}$ – time $t_p \in T$ transaction of information resource processing in the system $S$ when $p = \overline{1, n_T}$ ,
$Y = \{y_1, y_2 \ldots y_{n_Y}\}$ – set of statistical data $y_j \in Y$ in system at $j = \overline{1, n_Y}$ .

## 4. CONTENT FORMATION METHOD

The commercial content formation for information resource provides a link between the input data from different sources set and the commercial content set into the appropriate database in electronic content commerce systems that can be presented as $Sourse\langle i \rangle \rightarrow x_i \rightarrow X \rightarrow Formation(u_f, x_i, t_p) \rightarrow c_r \rightarrow C \rightarrow DataBase\langle C \rangle$, where $Sourse\langle i \rangle$ – content source, $x_i$ – matched content from the source, $X$ – the relevant sources data set, $Formation(u_f, x_i, t_p)$ – content formation operator in a fixed time $t_p$ under $u_f$ appropriate conditions, $c_r$ – formed content under $u_f$ conditions, $C$ – generated content set, $DataBase\langle C \rangle$ – commercial content prevailing database. Content formation model in electronic content commerce systems can be showed as

$$Formation = \left\langle \begin{matrix} X, Gathering, Formatting, KeyWords, Backup, \\ Caterization, BuDigest, Dissemination, T, C \end{matrix} \right\rangle, \qquad (2)$$

where: $X = \{x_1, x_2 \ldots, x_{n_X}\}$ – input data set $x_i \in X$ from different information resources or the moderators at $i = \overline{1, n_X}$ ,
*Gathering* – content collecting/creating operator from various sources;
*Formatting* – content formatting operator,
*KeyWords* – the content key words and concepts identify operator;
*Categorization* – content categorization operator,
*Backup* – the content duplicate detect operator,
*BuDigest* – content digest formation operator,
*Dissemination* – content selective distribution operator; $T = \{t_1, t_2 \ldots, t_{n_T}\}$ – the content forming transaction time $t_p \in T$ while $p = \overline{1, n_T}$ ,
$C = \{c_1, c_2 \ldots, c_{n_C}\}$ – a content set $c_r \in C$ with $r = \overline{1, n_C}$ .

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

$$c_r = \left\{ \bigcup_f u_f \middle| (x_i \in X) \wedge (\exists u_f \in U), U = U_{x_i} \vee U_{\overline{x_i}}, i = \overline{1,m},\ f = \overline{1,n} \right\} \qquad (3)$$

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 information from different sources, provide the data syndication $C = Gathering(X, U_G, T)$, i.e. information gathering from sources and its fragments for further distribution according to user needs, where $X$ – content set from data different sources, $U_G$ – data collecting conditions set from various sources, $Gathering$ – the content collecting/creating operator, $T$ – the content collection/creation time.

Content duplicate detecting marked by the operator as

$$C = Backup(Gathering(X, U_G, T), U_B), \qquad (4)$$

where: $X$ – content set from data different sources,

$U_B$ – text content duplication identify conditions set,

$Backup$ – the text content duplication identify operator,

$C$ – 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 (of information resources, from moderators, users, visitors, journalists and editors), content direct scanning and bringing the total:

$$C = Formatting(Backup(Gathering(X, U_G, T), U_B), U_{FR}), \qquad (5)$$

where: *Formatting* –content formatting operator,
$U_G$ – data collecting conditions set from various sources,
*Gathering* – the content collecting/creating operator,
$U_{FR}$ – information formatting conditions set,
$T$ – the content collection time.

A content set $C$ developing to keywords identify is built on the keywords finding principle in content (terms), based on the Zipf's law and reduced to the words choice with an occurrence average frequency (the words most used ignored by stop-dictionary, and rare words from messages text do not into account). Keywords and concepts identify defined by the operator *KeyWords* $\langle C \rangle$ and the operator described the form:

$$C = KeyWords\langle Formatting\langle Backup\langle Gathering(X, U_G, T), U_B\rangle, U_{FR}\rangle, U_K\rangle \quad (6)$$

,

where *KeyWords* – the content keywords and concepts identify operator that is implemented as a processes set, using the presented chart in Fig. 1; *Formatting* – content formatted operator; $U_G$ – data collecting conditions set from various sources; *Gathering* – the content collecting/creating operator; $U_{FR}$ – conditions data formatting set; $T$ – the content collection time; $U_K$ – keywords and concepts identify conditions set.

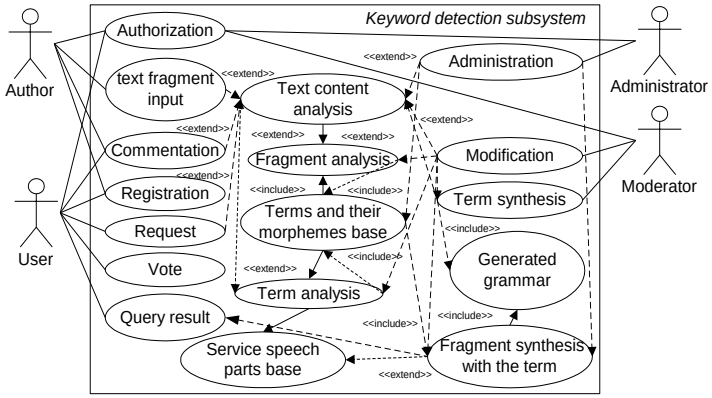


**Fig. 1. Use case diagram for the content keywords identifying process [source: own study]**

Content analysis for compliance thematic requests to $C_{Ct} = Categorization\langle KeyWords(C, U_K), U_{Ct}\rangle$, where $KeyWords(C, U_K)$ – the keywords identify operator, *Categorization* –content categorize operator according to

the keywords identified, $U_K$ – keywords identify conditions set, $U_{Ct}$ – categorization conditions set, $C_{Ct}$ – rubrics relevant content set. Digest set formed by such dependence as $C_D = BuDigest\langle C_{Ct}, U_D\rangle$, where $BuDigest$ – digests forming operator, $U_D$ – conditions set for the digests formation, $C_{Ct}$ – rubrics relevant content set, i.e.

$$C_D = BuDigest\langle Categorization\langle KeyWords(C, U_K), U_{Ct}\rangle, U_D\rangle. \quad (10)$$

Content sent by users and uploaded into thematic database. Content selective distribution described as $C_{Ds} = Dissemination\langle C_D, U_{Ds}\rangle$, where $C_{Ds}$ – content selectively distributed set, $U_{Ds}$ – content selective distribution conditions set, $Dissemination$ – the content selective distribution operator.

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).

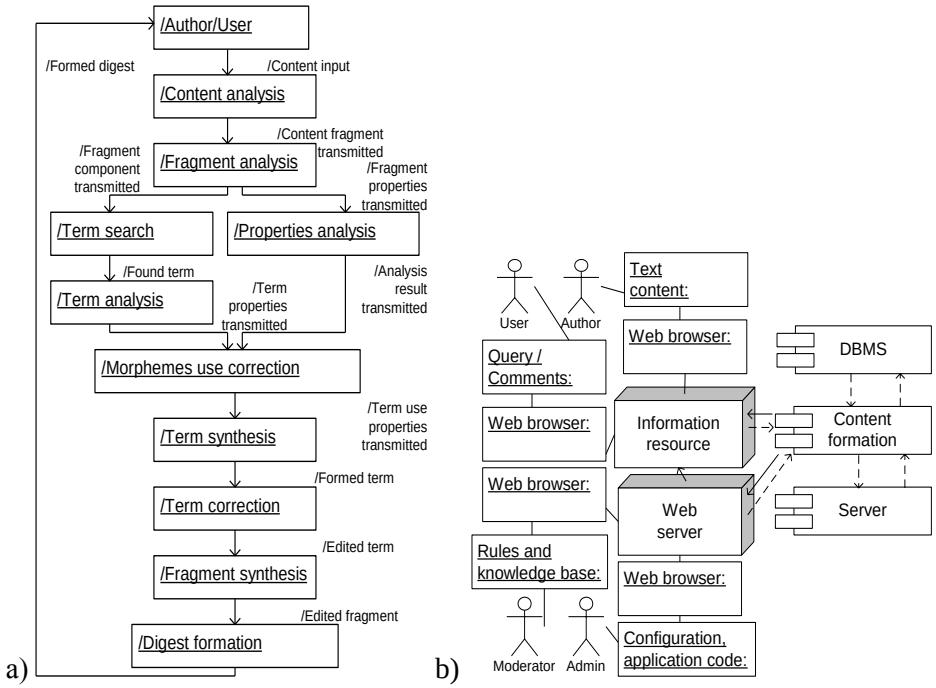


**Fig. 2. a) Cooperation diagram and b) components diagram for the process to content subject keywords find [source: own study]**

In Fig. 2, a submitted cooperation diagram for content subject keywords identifying process. The text lexical-grammatical and semantic-pragmatic construction analysis used in the 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 analysing is the new term synthesis as a content topic keyword. In Fig. 2, b submitted component diagram for content topic keyword process. The keywords detecting principle by content (terms) is based on the Zipf law. Process reduced to the words choice with use average frequency (the most-used words ignored by the dictionary, and rare words in the text do not include) using terms and their morphemes database. In Fig. 3 activity diagram for the content subject keywords identifying process is showed. The present method next step in the content forming is the content categorization.

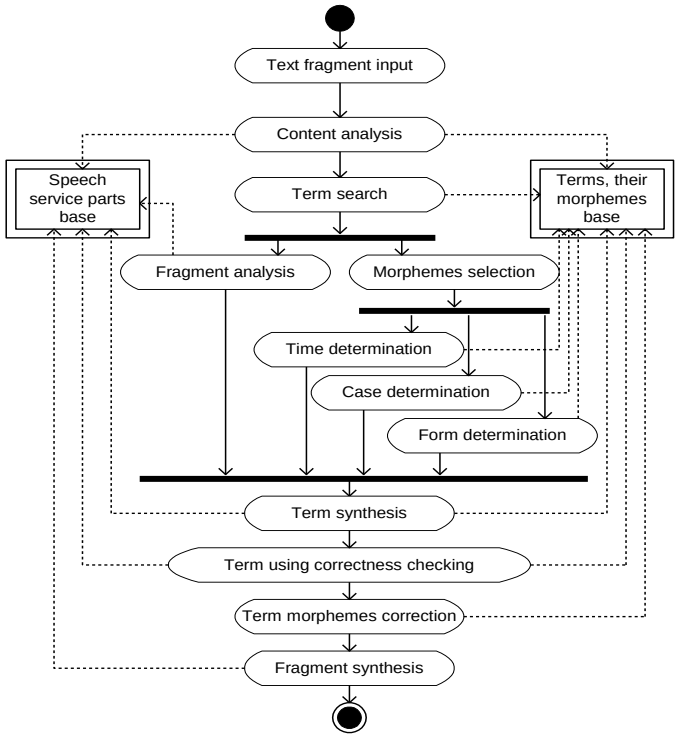


**Fig. 3. Activity diagram for the content subject keywords identifying process [source: own study]**

## 5. CONTENT MANAGEMENT METHOD

The commercial content life cycle presented in the form next major processes communication as *source* $\rightarrow$ *content formation* $\rightarrow$ *database* $\rightarrow$ *content management* $\rightarrow$ *user* $\rightarrow$ *content implementation* $\rightarrow$ *database.* ECCS model $S = \langle X, Formation, C, Management, Realization, Y \rangle$, where $X = x_1, x_2, \ldots, x_{n_X}$ – input data set, *Formation* – content formation operator, $C = c_1, c_2, \ldots, c_{n_C}$ – content set, *Management* – content management operator, *Realization* – content implementation operator, $Y = y_1, y_2, \ldots, y_{n_Y}$ – output data set. Below is the content management models classification.

1. Pages generate per request is submitted in the form of the following main stages connection *Content* $\rightarrow$ *content editing* $\rightarrow$ *Database* $\rightarrow$ *content presentation* $\rightarrow$ *informational resource.* Pages generate model on demand as $Management_Q = \langle X, C, Q, R, Edit, Y \rangle$, where $X$ – input data set, $C$ – content set, $Y$ – pages generated set, $Q$ – query set, $R$ – pages formulation and submission function, $Edit$ – content editing and updating function.
2. Pages generate model while editing is presented as the next major stage of communication *Content* $\rightarrow$ *content editing* $\rightarrow$ *database* $\rightarrow$ *informational resource.* When making changes to the site content creates a static *pages* set. Not is taken into account interactivity site between visitors and content. Pages generation system model while editing as $Management_E = \langle C, Edit, Y \rangle$, where $C$ – content set, $Y$ – static pages set, $Edit$ – content editing function. The pages formation described function as $\vec{y}\ t = Edit\ \vec{c}, Weight, t$ .
3. Pages generate mixed model combines the advantages of the first two types and is presented as a communications major stages *Content* $\rightarrow$ *content editing* $\rightarrow$ *Database* $\rightarrow$ *content analysis* $\rightarrow$ *blocks collection* $\rightarrow$ *content presentation* $\rightarrow$ *informational resource.* Mixed type model is as $Management_M = \langle X, C, Q, R, Edit, Caching, Y \rangle$, where $X$ – input data set, $C$ – content set, $Y$ – pages generated set, $Q$ – query set, $R$ – pages formulation and submission function, $Edit$ – content editing and updating function, *Caching* – cache formulation function. Cache is update automatically (after a certain period or when amending certain site sections) or manually (team administrator). Another approach is to maintain information blocks on the editing site stage. Then the pages collected from these units when requesting user. The process is implementing caching. The module generates submission page once. Then it is several times faster downloading from the cache.

User queries content analysis allows to qualitatively assess the content flow in the system. This facilitates the subsequent decisions by the moderator as follows: the problem situation description and study purpose search; precise definition of the study object and subject; the object preliminary analysis; concepts substantial clarification and empirical interpretation; the procedures description for the properties and phenomena registration; the overall study plan determining; the definition of the sample type, sources collection and so on. Qualitative content analysis is intended to provide the necessary means moderator for results analysis (Table 1). With their help identify the content properties and test them on the general content stream. Then apply the content stream general properties on its specific thematic part.

**Tab. 1. Qualitative content analysis stages**

| Stage name | Stage characteristics |
|---|---|
| Text wrapping on blocks | Integrated content units are formation for encoding and processing. |
| Content stream reconstruction | The values, thoughts, views systems and arguments are reconstruction of each source text. |
| Conclusions forming | Generalizations are withdrawal by comparing individual system values. |

Quantitative content analysis consists of the stages presented in Table 2. Main task for content management process are the following items: a database create and access to it; the operational and retrospective databases forming; databases rotation; users work personalization; personal needs and sources protection; work statistics keeping; search ensuring in database; output forms generation; interaction with databases of other subsystems. In Table 3 presents the content management key stages in the electronic content commerce systems.

**Tab. 2. Quantitative content analysis stages**

| Stage name | Stage characteristics |
|---|---|
| The analysis unit selecting | Linguistic unit convert in the form for processing. |
| Units frequency counting | Relationships are identifying between linguistic units. |
| Categorization | Categories finite and excess aggregate are determining to obtain quantitative data of their appearance. Categories irregular sequence is clustering (into groups and classes division). And on the basis of new generalized categories is received. |
| Data Mining | New knowledge is identifying in the content flow through multiple quantitative evaluations. Next qualify them as categories. |
| Results interpretation | Content and semantically-filled results are getting. For this purpose use various statistical mathematical methods and semantic formalisms. |

**Tab. 3. Content management stages**

| Stage name | Process name | Process features |
|---|---|---|
| Content editing | content themes definition | creating goal, the content and structure formation; |
| | form definition of content presentation | graphic information; the text (article, news release, job descriptions); HTML templates; back-end code, etc.; |
| | management tools selection | HTML editors, processors word; visual tools for creating objects; |
| Content analysis | rights access assignment | full or limited access to content; |
| | process identifying | standard processes of new information content creation/ publishing; |
| | content saving | in a database or repository; |
| | processes logging | creation, transmission and storage processes; |
| | information interactive | information about the next performer content; |
| | events audit | content versions save; |
| | text content analysis | quantitative or qualitative; |
| | versions access | support the possibility of users appealing to previous content versions; |
| | business process analysis | objectives, roles and tasks definition; roles default to user groups; business processes development for all content; |
| Content presentation | static | without any logic behavior; |
| | dynamic | personalization (rules/filters), globalization, localization. |

As information technology basis considered annotated database in search engines. It contains an index, inverse, dictionary tables, etc. In content management systems creates a database search primary content pattern (PCP). They used clustering technology (automatic forming groups with similar content on the criteria PCP). In content management systems formed database annotations for used in the search process. Clusters database each record is corresponding cluster definitions and containing its description. Database record is performing automatic abstracting methods (digest is formation of text statistical analysis methods). These methods are used to create the PCP and descriptions of available users (Fig. 4). Personalization based rules is the content provision to specific users or users groups of conventional business logic using. For example, using a rule where all those interested in children's books fall into the group that focused advertising children's clothing. Rules developed on the content basis that type users in a registration card. In content management systems use algorithms categorization with personalization using filters (intelligent agents). Also, algorithms use based on the content analysis of user behaviour. In particular, he analyses the content to which the user accesses, the sites visited and more. That is constantly conducted analysed registered user and user’s group history with priorities an overwhelming number for interest.

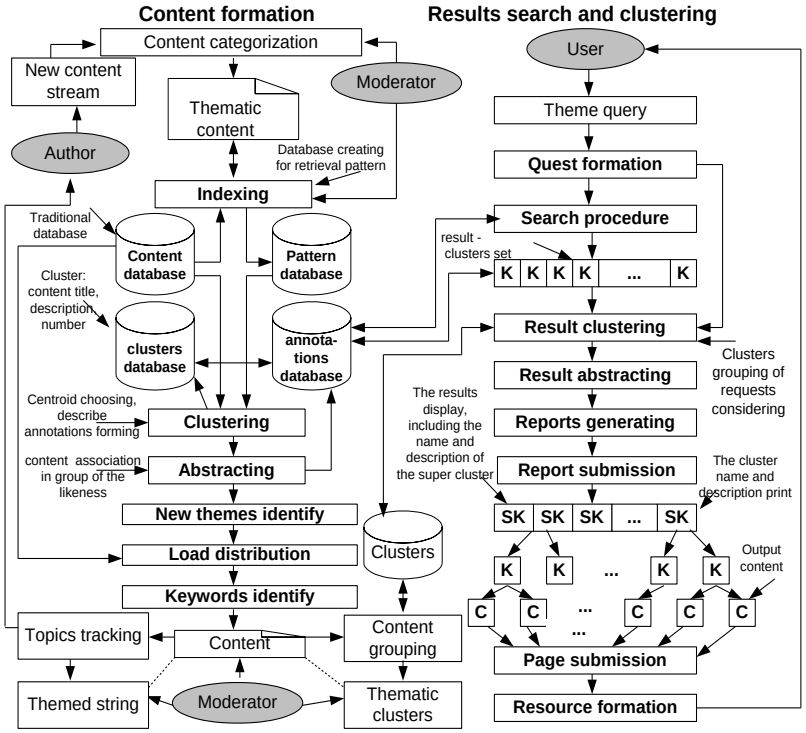


**Fig. 4. Scheme of the electronic content commerce systems functioning from annotated database [source: own study]**

The full-text search problem in large content arrays is ineffective. The annotated content search solves the problem exactly: instead of the full content searching to search on annotations (search content pattern). Digest remotely similar content and often not perceived by person. But as the search content pattern with weighted keywords and phrases it leads to adequate results with full-text search. Digest constructed from content fragments with the largest weight values. Content analysis is used for digests automatic generation, the concepts (categories) relationship automatic detection, relationships automatic clustering to the most important identify, the relationships automatic detection (e.g., positive and negative). One of the most important tasks in content analysis is the categorization process. It sets the conceptual grid. In its terms is the content flow analysing and new categories generate.

## 6. CONTENT SUPPORT METHOD

Content support method provides information portraits formation; digests formation; thematic content identification; concepts relationship tables' construction; concepts ratings calculation; new developments identification,

their tracking and clustering. The fully functional ECCS is characterized complex system of interrelated operations, methods, techniques. The annotations database creating is a search images database creating of original content and their clustering (to the content group formation with close on some criteria of search images). Annotation base (search pattern of clusters used in the search process) is associated with the cluster base. Each entry corresponds to its specified cluster and description containing. This description is made of automatic abstracting methods.

Content analysis of textual information allows determine the signs prevalence of researched content in ECCS. It is important not as an absolute, but the relative importance of attributes, i.e. characteristics of the place (shares) among other features. For example, this is the percentage of forum users discussion of economic Issues concerning political or the percentage of positive comments on articles regarding the negative and in respect of all comments on this category of articles in online newspapers. The degree of correlation between the features in the texts provides empirical data to understand of the functional links between elements reflected in the texts reality e.g. the audience mood determination of online newspaper about the economic or political situation in the country and/or world. If there are texts that have chronological sequence received number fixed in time portraits of the investigated reality (change in demand for a content category according to the season, such as winter read more fiction and detective stories – summer) or the target audience portraits (change in demand for a content category according to the article, for example, demand for political articles before the election). This allows to put forward hypotheses of the prognostic character about the system functioning.

In digest formation using content analysis with regard to frequency weights of words from the concepts dictionary generated. The digest formation consists of algorithms of the concepts dictionary forming (alg. 1), of the content duplicate definition (alg. 2) and of the digest create (alg. 3).

Algorithm 1. Concepts thematic dictionary formation.

**Stage 1.** Concepts alphabetical-frequency vocabulary formation.

*Step 1.* Sequential selection of all words in the input content stream.

*Step 2.* The alphabetical-frequency vocabulary construction based on content categories.

*Step 3.* Words normalization through automatic morphological analysis.

*Step 4.* The alphabetical-frequency vocabulary modification.

*Step 5.* Words assign of weight $w$ (use frequency).

*Step 6.* Insignificant words deleting from the alphabetical-frequency vocabulary ($W \le k$, where $k$ – threshold value of word extracting).

**Stage 2.** The thematic dictionary choice as requested.

**Stage 3.** Words weight adjustment of alphabetical-frequency vocabulary dictionary based on thematic dictionary.

**Stage 4.** Words choosing at $N = n$ are with more weight $w$ of alphabetical-frequency vocabulary, which $n = const$ is given by the moderator.

Algorithm 2. Duplicate content determination.

**Stage 1.** The initial data formation.

*Step 1.* The moderator introduced of words in a string $m = const$.

*Step 2.* The moderator input strings unique coefficient ie $U = const$.

*Step 3.* Coefficient limits formation $K = \{a_1, a_2\}$ of keywords use, where $a_1 = const$, $a_2 = const$.

*Step 4.* Content partitioning on $n$ chains of $m$ words.

*Step 5.* Frequency calculation of $k_i$ keywords use.

**Stage 2.** Content duplicates determination.

*Step 1.* Words strings comparison for all content.

*Step 2.* Chains uniqueness coefficients calculation $u_i$.

*Step 3.* Chains uniqueness coefficients comparison $u_i$ with $U$. At $\frac{1}{n}\sum_{i}^{n} u_i < U$ mark the content as unsuitable.

*Step 4.* The frequency comparison $k_i$ with coefficient $K$. If $k_i < a_1$ or $k_i > a_{21}$ then a content mark as unsuitable.

Algorithm 3. A digest create.

**Stage 1.** Content Select based on its weight.

*Step 1.* Digest size $C$ input.

*Step 2.* The algorithm 1 implementation.

*Step 3.* Weight consistent determine of each content as $W = \sum w_i$ sum of its individual words.

*Step 4.* The input content stream sort from the weights values.

*Step 5.* Meaningful content duplicates definition for statistical criterion of text uniqueness $U \geq 0{,}9$ (alg. 2).

*Step 6.* Content filter of unsuitable for digests building (when $W \leq l$, where $l$ – Content removal threshold value by the self-education rules of content structuring and moderating) and statistically substantial duplicates.

*Step 7.* Choice of $V = q$ content with greater weight where $q = const$ and the moderator given.

**Stage 2.** Digest text construction of selected content.

*Step 1.* Dictionary construction of selected content (alg. 1).

*Step 2.* Content analysis application to the text (Table 4).

*Step 3.* Sentences filtration that do not meet semantic rules of content structuring and moderating.

*Step 4.* Hypertext presentation formation of digest, its contents and a link to the original source.

**Stage 3.** Generated text edit of digest.

*Step 1.* The check amount of generated content $c_i$. If $c_i < C$, then step 2, otherwise stage 4.

*Step 2.* Content delete from the input stream that is used to the digest formation.

*Step 3.* Steps 1-2 implementation.

*Step 4.* Resulting append to the pre-formed digest and move to step 1.

**Stage 4.** Digest text formation as a separate content and its maintaining in the database with reference on the source.

**Tab. 4. Content analysis stages of textual information**

| Stage | Stage characteristics of content analysis |
|---|---|
| Total sources or content determination | Using a set of defined criteria which corresponds to each content: given type of source; one type of content; given the parties which involved in the communication process; message size matched (minimum / length); messages appearance frequency; messages distributing method; messages distribution space; messages appearance time, etc. |
| Content analysis selection | The selected set of content is formed on the criteria a limited sampling from a larger array of information. It's forming using the procedure from a set of precisely defined actions for processing without any changes of all objects study. |
| Linguistic units identifying | Compliance with strict requirements concerning the linguistic units choice for content analysis: large enough to interpret meaning; small enough not to interpret the many meanings; easily identified; units number is large enough for sampling. When taking of the themes analysis unit take into account that its size does not go beyond a paragraph; new theme arises with the new characteristics appearance of units. |
| Computation units finding and classifier formation | Computing units may coincide with semantic units, or have specific characteristics. In the first case, the analysis procedure is to the frequency calculate of the selected content unit use. Otherwise, the researcher proposes computation unit (physical length of the texts; text area, filled with informative units; the number of rows, paragraphs, characters, columns in text; the file size / type; pictures number with a certain content and story) based on the analysed material and research purposes. |
| Calculating | Standard techniques for classification of selected groups of mathematical statistics and probability theory formulas. |
| Results interpretation | This includes all extracted text fragments. When forming conclusions do not take into account of the some results, without exception all. Here are identified and measured the text characteristics. They allow drawing conclusions about that wanted to emphasize or hide its author. Or they predict changes in demand for content based statistical set of calculated coefficients for the time period of specified category. |

**Topical stories identification in content stream.** Content with new themes is the new group's basis of interdependent content in thematic stories identifying with the following procedures: control within the system – level destination of user access to different content; content integration – content moving to a new decision; content support of various types – content storage and sorting in a central repository; detailed documentation and context-intelligent help support; rating system of site articles evaluation; template changes – general formatting changes to the content of the part site reflects the entire site; workflow support – automated business processes create for specific content; content marking – new categories and markers adding to content before / after saving; version control –new versions creating, view and return to the previous versions of content; content analysis of text streams in the system; visual administration tool – easy authors management of content, without resorting to programming, typically implemented using HTML-forms; concepts relationship tables construction.

**The concepts ratings calculation** is based on procedure for results calculating of content analysis, taking into account the ratio coefficient *c* of positive and negative (for the selected item) estimates, opinions, arguments, as described in the user comments on the content ECCS.

## 7. ANALYSIS OF DATA

Based on the developed method content forming subsystems at various stages are implemented in Internet projects “Fotoghalereja Vysocjkykh” (FGV, fotoghalereja-vysocjkykh.com), “Tatyana” (T, tatjana.in.ua), “AutoChip” (autochip.vn.ua), “Vgolos” (vgolos.com.ua), “PressTime” (presstime.com.ua), “Exchange rates” (ER, kursyvalyut.com), “Good morning, accountant!” (GM, dobryjranok.com). Table 5 presents the developed systems comparative characteristics derived from Google Analytics.

**Tab. 5. The system work comparative characteristics for the time period 10.2012-11.2012**

| Systems characte rization | FGV | Vgolos | T | PressTime | Auto Chip | ER | GM |
|---|---|---|---|---|---|---|---|
| Formation | +/- | + | - | +/- | + | + | +/- |
| Managment | + | + | - | + | + | +/- | + |
| Support | +/- | + | - | + | +/- | +/- | + |
| Visiting | 73 | 326940 | 49 | 167 856 | 406 | 103 | 58 |
| Unique visitors | 62 | 217719 | 21 | 123 756 | 326 | 42 | 7 |
| Pages browse | 136 | 562455 | 142 | 345 234 | 863 | 237 | 226 |
| Pages/Visit | 1,86 | 1,72 | 2,90 | 1,45 | 2,13 | 1,67 | 3,90 |
| Visits average duration | 00:47 | 01:45 | 04:38 | 01:09 | 01:08 | 00:37 | 09:35 |
| Fault indicator | 71,23 | 76,92 | 46,94 | 79,56 | 56,90 | 61,23 | 48,28 |
| New Visits | 80,82 | 51,83 | 36,43 | 45,65 | 77,59 | 90,87 | 12,07 |
| Returning Visitor | 82,19 | 48,15 | 63,27 | 54,35 | 77,59 | 62,79 | 87,93 |
| New Visitor | 17,81 | 51,85 | 36,43 | 45,65 | 22,41 | 37,21 | 12,07 |
| Ukraine | 87,67 | 89,81 | 71,43 | 92,33 | 73,89 | 97,07 | 55,17 |
| Russia | 2,74 | 2,55 | 24,49 | 6,27 | 17,00 | 1,05 | 43,10 |
| USA | 1,37 | 0,58 | 0,07 | 0,06 | 0,05 | 0,61 | 1,72 |
| Search traffic | 69,86 | 36,03 | 73,47 | 60,05 | 88,67 | 59,03 | 43,10 |
| Traffic Conversion | 12,33 | 54,62 | 0 | 34,65 | 3,45 | 35,65 | 6,90 |
| Direct traffic | 17,82 | 9,21 | 26,53 | 5,25 | 7,88 | 5,32 | 50,00 |
| Traffic campaigns | 0 | 0,14 | 0 | 0,05 | 0 | 0 | 0 |

The commercial 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 commercial 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 – analysed text partition, sentences and tokens table. There is the following relationship: the more unique content in the ECCS,

the more the visitor's information resource in its system. Google Analytics provides advanced data analysis and allows us to estimate the content traffic and marketing activities effectiveness, such as the newspaper "Vgolos" (Fig. 5).

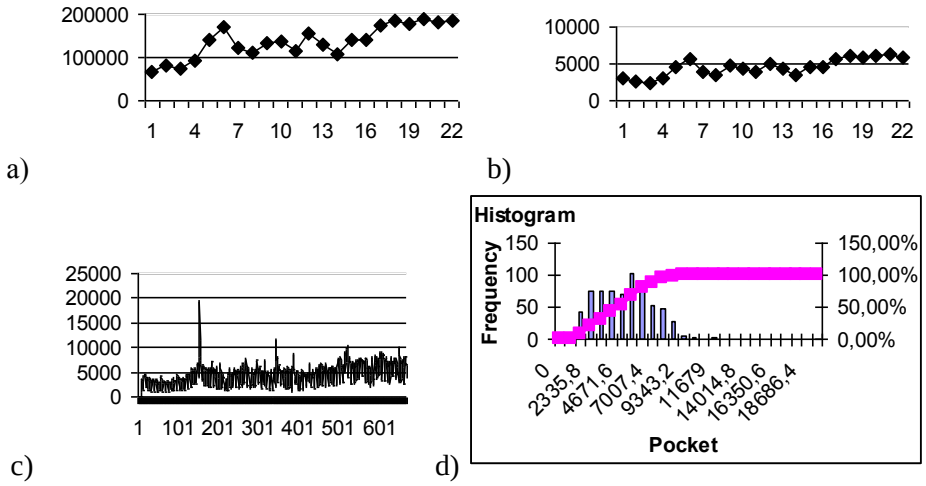


**Fig. 5. Visitors distribution of a) total, b) medium, c) daily, d) monthly number in 2010-2012 [source: own study]**

## 8. CONCLUSIONS

In the given paper is functional logistic model of commercial 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; key words and concepts identifying; content categorization; content duplicate detection; digests formation; selective distribution of content between moderators and users of ECCS. 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 commercial content projects. Developed automation commercial content processing allows speeding up the content formation, management and maintenance process. It also contributes to the rating increase of generated by their use with commercial information resources.

**REFERENCES**


[1] BEREZA A., KOZAK I., LEVCHENKO F.: *Elektronna komertsiya*. (in Ukraine). Kyiv: KNEU, 2002, p. 326.

[2] BERKO A., VYSOTSKA V., PASICHNYK V.: *Systemy elektronnoyi kontent-komertsiyi*. (in Ukraine). Lviv: NULP, 2009, p. 612.

[3] BOLSHAKOVA E. et al.: *Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i kompyuternaya lingvistika.* (in Russian). Moskva: MIEM 2011, p. 272.

[4. BRAYCHEVSKY S., LANDE D.: *Sovremennye informatsionnye potoki*. (in Ukraine). Nauchno-tehnicheskaya informatsiya, 2005, № 11, pp. 21-33.

[5] CLIFTON B.: *Google Analytics: professionalny analiz poseschaemosti web-saytov.* (in Russian). Moskva: Vilyams, 2009, p. 400.

[6] KORNEEV V. et al.: *Bazy dannyh. Intellektualnaya obrabotka informatsii.* (in Russian). Moskva: Nolidzh, 2000, p.352.

[7] LANDE D. et al.: *Osnovy modelirovaniya i otsenki elektronnyh informatsionnyh potokov*. (in Ukraine). Kyiv: Inzhiniring, 2006, p. 348.

[8] LANDE D.: *Fundamentals information streams integration*. (in Ukraine). Kyiv: Engineering Publ., 2006, p. 240.

[9] PASICHNYK V. et al.: *Matematychna linhvistyka.* (in Ukraine). Lviv: "Novyy Svit – 2000", 2012, p.359.

[10] SOVETOV B., YAKOVLEV S.: *Systems modeling.* (in Russian). Moskva: VS Publ., 2001, p. 343.

[11] FEDORCHUK A.: *Content Monitoring of information streams*. (in Ukraine). Nat. Acad. Science Library № 3., Kiev, 2005.

[12] KHARCHENKO K.: *Text information computing content analysis* [Online]. Available: http://leveltar.narod.ru/ cta/teor.html. [Accessed: Dec. 9, 2013].

[13] BOIKO B.: *Content Management Bible.* (in USA). Hoboken, 2004, p. 1176.

[14] CM LIFECYCLE POSTER: *Content Management Professionals*. (in USA). Retrieved 20 July 2010. http://www.cmprosold.org/resources/poster/.

[15] DOYLE B. *Seven Stages of the CM Lifecycle*. (in USA). EcontentMag.com. 2010. www.econtentmag.com/Articles/ArticleReader.aspx?ArticleID=13554&AuthorID=155.

[16] HACKOS J. *Content Management for Dynamic Web Delivery*. (in USA). Hoboken, NJ: Wiley, 2002, p.432.

[17] HALVERSON K.: *Content Strategy for the Web.* (in USA). Reading, Mass: New Riders Press, 2009. – 192 p.

[18] MCGOVERN G., NORTON R.: *Content Critical*. (in USA). Upper Saddle River, NJ: FT Press, 2001, p. 256.

[19] MCKEEVER S.: *Understanding Web content management systems: evolution, lifecycle and market*. (in USA). Industrial Management & Data Systems (MCB UP), 2003, 103 (9), pp. 686-692.

[20. NAKANO R.: *Web content management: a collaborative approach*. (in USA). Boston: Addison Wesley Professional, 2002, p. 222.

[21] PAPKA R.: *On-line News Event Detection, Clustering, and Tracking*. (in USA). Ph. D. Thesis, University of Massachusetts at Amherst, September, 1999.

[22] WOODS R.: *Defining a Model for Content Governance*, 2010, www.contentmanager.net/ magazine/article_785_defining_a_model_for_content_governance.html.

[23] ROCKLEY A.: *Managing Enterprise Content: A Unified Content Strategy*. (in USA). Reading, Mass: New Riders Press, 2002. – 592 p.