Content Formation Method in the Electronic Content Commerce Systems

Andriy Berko[1], Victoria. Vysotska[2], Lyubomyr Chyrun[3]

[1]Environment and Ecosystems Department, Lviv Polytechnic National University, UKRAINE, Lviv, S. Bandery street 12

[2]Information Systems and Networks Department, Lviv Polytechnic National University, UKRAINE, Lviv, S. Bandery street 12, E-mail: victana@bk.ru

[3]Software Department, Lviv Polytechnic National University, UKRAINE, Lviv, S. Bandery street 12

Abstract – In the given article content is forming method as the content life cycle stage in electronic commerce systems is proposed. The method implements the information resources processing in electronic content commerce systems and automation technology simplifies the commercial content formation. In the given article the main problems of electronic content commerce and functional services of commercial content forming are analyzed. The proposed method gives an opportunity to create an instrument of information resources processing and to implement the module of content forming.

Key words – information resources, commercial content, content analysis, content monitoring, content search, electronic content commerce systems.

I. Introduction

The Internet active development promotes the needs growth in production/strategic data and new forms of information services implementation [6, 9]. Documented information is an informational product or commercial content, if it is prepared in accordance with user needs and intended to meet them. 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 [6].

II. Recent research and publications analysis

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 the its spread speed. In addition there was also a specific problems number (Table 1) [9].

TABLE 1

THE COMMERCIAL 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 content streams/sources. Their application caused needs by the systematic tracking of trends and processes in the content environment that is constantly updated. 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 [8, 9, 12, 14]. 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 words, phrases or sentences.

Great importance is the 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 the test 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 the presence desired linguistic unit in 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 text content is served in terms of specialized information retrieval language [8-14]. SCP determination procedure is the main content text indexing, semantic analysis and translating it into information retrieval language (Table 2) [8, 12, 14]. The module does not retain content text and it's SCP. To search of indexed content used content analysis to information 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 content indexing [12-16]. Content-search contains a semantic tools set: information retrieval language, content/queries indexing and search methods [12-16]. The semantic tools basis is information retrieval language. This is specialized artificial language, which is intended to describe the central content themes/subjects and formal characteristics, as well as to describe the requests and searching capabilities [12-16]. In practice, one language is used for indexing content, and another – for information requests indexing.

TABLE 2

THE MAIN STAGES IN CONTENT-SEARCH OPERATION

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 content analysis result is positive in the range (0.7, 1] or (0.5, 1].
Decision-making	The decision on issuing the content according to the applying content analysis result.
Content submission	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 content performed the moderator manually or automatically by content analysis means [9]. During indexing explore content text, determine its central theme and describe it in terms of information retrieval language [12-16]. In the section names content usually reveal a central theme and subject, but the name is not always possible to the content identify. Natural language is not used as an information retrieval language through numerous grammatical inclusions, structuring lack, ambiguity and greater redundancy, particularly 75-80% for the Ukrainian language. 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.
Grammar (syntagmatic)	The rules set for combining linguistic units in phrases, i.e. sentences constructing effective means.
Paradigms	Words lexical-semantic group with a subject-logical relations based on semantic features.
Paradigmatic relation	Relation basic and analytical between words that do not depend on the context in which they are used, and connections caused not linguistic and logical.
Syntagmatic relation	Linear relation between words, which are set when words combining into phrases and phrases.
Indexes identification rules	This is a language paradigmatic (vocabulary) and syntagmatic (grammar).
Sentences	Implemented sentence, i.e. statement is a sentence in natural language, but the reverse is not true.
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.
Fragments blocks	This is a between phrase unity set that provide integrity through text content and thematic connections.

The using feasibility of information retrieval languages depends on the search tools purpose, the technical equipment level, information procedures and management level automation [12-16]. 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 electronic content commerce systems [12-13].

III. Problems selection

Text content (articles, comments, book, etc.) contains a significant amount of text in natural language, where the information is abstract. Text is united for the content lexical items sequence, the basic properties which are informational, structural, communicative coherence and integrity. They reflect the content and structural essence of the text [3]. As functional-semantic structural unity of the text has construction rules, discovers patterns and formal connection meaningful of constituent units [2-3, 13]. Text connectivity is determined by external structural

parameters and formal dependence of text components, and integrity – through thematic, conceptual and modal dependence [3]. Text implements structural submitted activity that involves subject and object, process, purpose, means and results. They are displayed in the content-structural, functional, communicative performance of text. The text semantics due communicative task of content transfer [13]. The outside (composition) and the internal structure of text 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.

IV. Goals formulation

The article purpose is to develop a commercial content forming method for information resources processing in electronic content commerce systems. The work relevance is the need to obtain operational/ objectively assess of the competition level in the financial market segment of commercial content; assess the competition level and the competitiveness degree in the financial market with content distribution. From the systematic approach standpoint to investigate stages of the information resources processing and optimal life cycle develop for the content formation. The method development of content forming is enables the means of information resources processing and automatic generation of commercial content.

V. Research results analysis

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 $Sourse(x_i) \rightarrow x_i \rightarrow X \rightarrow Formation(u_f, x_i, t_p) \rightarrow c_r \rightarrow C \rightarrow DataBase(C)$, where $Sourse(x_i)$ – 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, $DataBase(C)$ – commercial content prevailing database.

Content formation model in electronic content commerce 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 $X = \{x_1, x_2 \ldots, x_{n_X}\}$ – 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:

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

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 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 $C = Backup(Gathering(X, U_G, T), U_B)$, 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:

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

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 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 $KeyWords\left(C\right)$ and the operator described the form:

$$C = KeyWords\left(Formatting\left(Backup\left(Gathering\left(X, U_G, T\right), U_B\right), U_{FR}\right), U_K\right)$$

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.

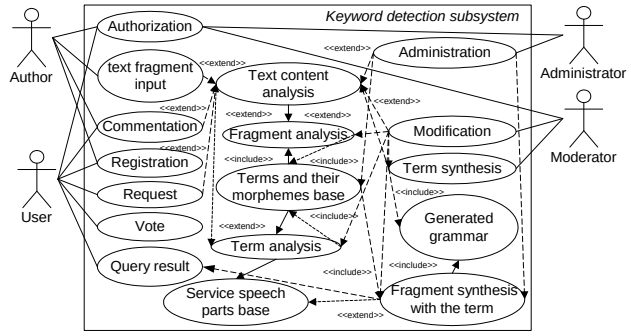

Fig. 1. Use case diagram for the content keywords identifying process

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

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

Content sent by users and uploaded into thematic database. Content selective distribution described as $C_{Ds} = Dissemination\left(C_D, U_{Ds}\right)$, where C_{Ds} – content selectively distributed set, U_{Ds} – content selective distribution conditions set, *Dissemination* – the content selective distribution operator. In Fig. 2, a submitted cooperation diagram for content subject keywords identifying process.

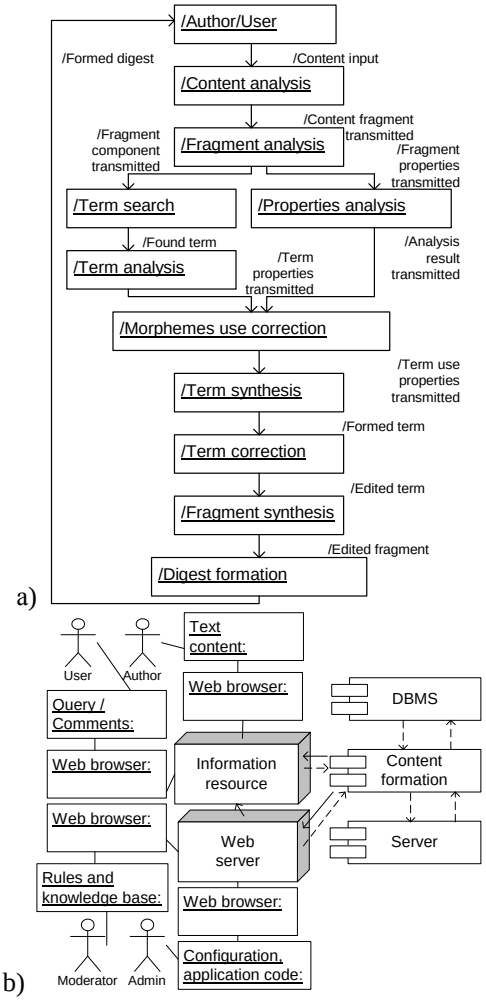

Fig. 2. a) Cooperation diagram and b) components diagram for the process to content subject keywords find

The text lexical-grammatical and semantic-pragmatic construction analysis used in 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 analyzing 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.

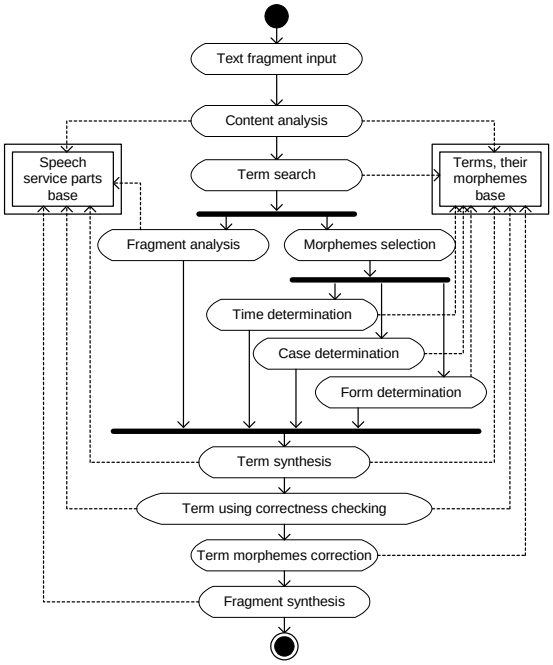

Fig. 3. Activity diagram for the content subject key words identifying process

Based on the developed method content forming subsystems at various stages are implemented in Internet projects "Fotoghalereja Vysocjkykh" (fotoghalereja-vysocjkykh.com), "Tatyana" (tatjana.in.ua), "AutoChip" (autochip.vn.ua), "Vgolos" (vgolos.com.ua), "PressTime" (presstime.com.ua), "Exchange rates" (kursyvalyut.com), "Good morning, accountant!" (dobryjranok.com). 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 11.2012 YEARS

Systems characterization	Fotoghalereja	Vgolos	Tatyana	PressTime	AutoChip	Exchange rates	Good morning
Content formation	+/-	+	-	+/-	+	+	+/-
Visiting	73	326 940	49	167 856	406	103	58
Unique visitors	62	217 719	21	123 756	326	42	7
Pages browse	136	562 455	142	345 234	863	237	226

Systems characterization	Fotoghalereja	Vgolos	Tatyana	PressTime	AutoChip	Exchange rates	Good morning
Pages/Visit	1,86	1,72	2,90	1,45	2,13	1,67	3,90
The visits average duration (min: c)	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
Visitors from Ukraine (%)	87,67	89,81	71,43	92,33	73,89	97,07	55,17
Visitors from Russia (%)	2,74	2,55	24,49	6,27	17,00	1,05	43,10
Visitors from the United States (%)	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

Google Analytics provides advanced data analysis and allows us to estimate the content traffic and marketing activities effectiveness, such as the newspaper "Vgolos" (Fig. 4).

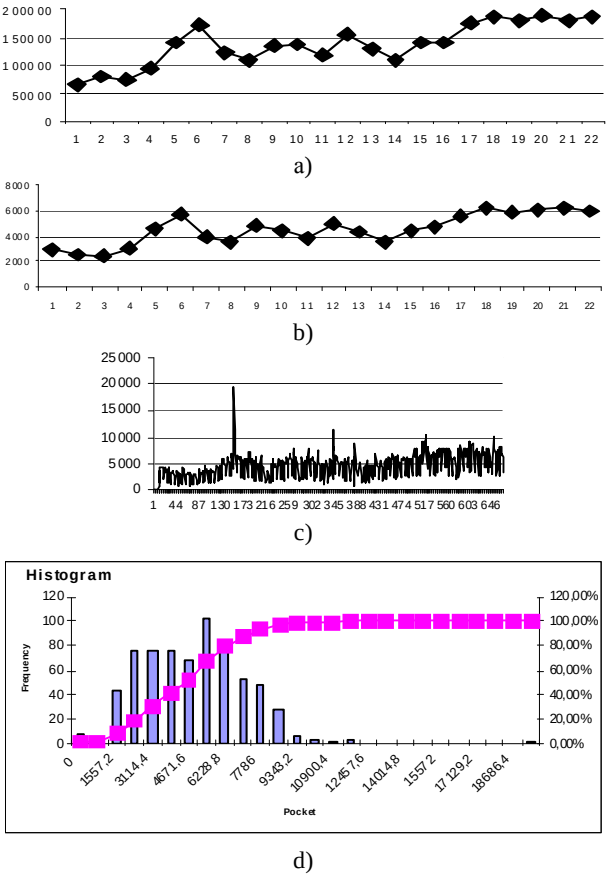

Fig. 4. Visitors distribution of a) total, b) medium and c) daily, d) monthly number in 2010-2012 years

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 – analyzed text partition, sentences and tokens table. There is the following relationship: the more unique content in the electronic content commerce systems, the more the visitor's information resource in its system. Commercial 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 electronic content commerce systems architecture.

Conclusion

The given article describes a commercial 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 electronic content commerce 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 commercial 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 commercial information resources.

References

[1] L. Averyanov, Content analysis, 2009 [Online]. Available: http://www.sbiblio.com/biblio/ archive/ averjanov_kontent/. [Accessed: Dec. 9, 2012].

[2] E. Bolshakova, D. Lande, A Noskov,. E. Klyshynsky, O. Peskova, E. Yahunova, Naturally texts automatic processing on language and computer linguistic, M.: MYEM Publ., 272 p., 2011.

[3] N. Valhyna, Theory text, Manual, M:Logos, 280 p., 2003 [Online]. Available: http://evartist.narod.ru/text 14/01.htm. [Accessed: Dec. 9, 2012].

[4] S. Grigoriev, Content analysis holding, [Online]. Available: http://www.psyfactor.org/lib/k-a2.htm. [Accessed: Dec. 9, 2012].

[5] I. Dmitriev, Content analysis: essence, problems, treatments, Online]. Available: http://www. psyfactor.org/lib/ka.htm. [Accessed: Dec. 9, 2012].

[6] B. Clifton, Google Analytics: professional attendance analysis web sites, M: Williams Publ., 400 p., 2009.

[7] G. Lelikov, V. Soroko, A. Grigoriev, D. Lande, Executive power monitoring with the electronic media computer systems content analysis use, Ukraine Civil Service Bulletin, N 2 '2002 [Online]. Available: http://www.infostream.ua/infostream/ publ/guds/ index.shtml. [Accessed: Dec. 9, 2012].

[8] D. Lande, A. Snarskyy, Y. Bezsudnov, Internetica. Navigation in complex net: models and algorithms, Moscow-2009 [Online]. Available: http://webgr ound.su/services.php?param=book&part=internetica_ content.htm. [Accessed: Dec. 9, 2012].

[9] D. Lande, V. Furashev, S. Braychevsky, O. Grigoriev, Modeling and evaluation fundamentals of electronic information streams, K.: Engineering, 348 p., 2006.

[10] D. Lande, A. Litvin, Phenomenon modern information flows, Journal "Net and Business", N 1, 2001 [Online]. Available: http://www.infostream.ua/ publ/ content/. [Accessed: Dec. 9, 2012].

[11] O. Manaev, Content analysis is the method description. Content analysis as research method [Online]. Available: http://www.psyfactor.org/lib/ kontent.htm. [Accessed: Dec. 9, 2012].

[12] T. Pisarevska, Information Systems and Technology in Human Resource Management, K.: MBK Publ., 279 p., 2000.

[13] M. Popov, A. Zaboleeva-Zotova, S. Fomenkov, Visualization semantic structure and abstracting texts on Naturally language, VTU [Online]. Available: http://www.dialog-21.ru/Archive/2003/Popov.htm. [Accessed: Dec. 9, 2012].

[14] V. Sitnik, T. Pisarevska, N. Eremin, O. Rrayeva, Fundamentals of Information Systems, K.: MBK Publ., 420 p., 2001.

[15] A. Fedorchuk, Content Monitoring of information streams, Nat. Acad. Science Library, Kiev, 2005.

[16] K. Kharchenko, Text information computing content analysis [Online]. Available: http://leveltar.narod.ru/ cta/teor.html. [Accessed: Dec. 9, 2012].