# Web Content Processing Method for Electronic Business Systems

Victoria Vysotska, Lyubomyr Chyrun

Information Systems and Networks Department, Lviv Polytechnic National University,

S. Bandery Str., 12, Lviv, 79013, UKRAINE

victana@bk.ru

Software Department, Lviv Polytechnic National University,

S. Bandery Str., 12, Lviv, 79013, UKRAINE

chyrunlv@mail.ru

## ABSTRACT

The given article is devoted to the development of standardized methods and software Web content processing in e-business systems. The model of electronic content commerce systems is developed. The models of commercial Web content processing subsystems are constructed. A new approach of business processes application and implementation for the construction of electronic content commerce systems is formulated. Complex methods of formation, management and support commercial content are developed. Software for information resources processing in electronic content commerce systems is developed. The methods of e-business systems designing and implementation are presented as example on online newspaper and online magazine that reflect the theoretical studies results.

## Indexing terms/Keywords

Web content; commercial conten; information resource; business process; online newspaper; online magazine;

 e-commerce systems; e-business.

## Academic Discipline And Sub-Disciplines

Computer science; systems analysis; web development; e-business

## SUBJECT CLASSIFICATION

Software products programming and creation; e-business systems design

## TYPE (METHOD/APPROACH)

Theme actuality consists in rapid prevalence of Internet access; active development of e-business; information products/services set expanding; demand for information products/services growing; theoretical justification methods lack of information resources processing; needed to software unification for information resources processing and active development of research in e-business of Google, AIIM, CM Professionals organization, EMC, IBM, Microsoft Alfresco, Open Text, Oracle, SAP corporations and McKeever S., Boiko B., McGovern G., Hackos J., Rockley A., Nakano R., Doyle B., Woods R., Halverson [8-21] in scientific papers.

## INTRODUCTION

Market of content distribution provides for the process of operational content preparation. It is available through information resources. Also, it is dependent on its values perception, reflection and preserving. Information systems moderators formalize, analyze, format and structure the content for a problems number processing and solving. The structuring process is definition of content pieces, methods and procedures of their combination with each other. So it is larger elements formation of the content with the smaller [1, 5]. The resulting content is stored in the database and/or data warehouse. There subsystem defines its areas/topics, such as electronic publications with higher coefficient of demand for the information resource visitors/users. Structured content is focused, such in ERP/CRM. But unstructured content is presented in an email, working free-form documents and teamwork instruments and stored, for example, in ECMS [1].

Web content lifecycle is a complex process. Content passes in it during the management through various stages/phases of the publication with a properties set (collaboration, records management, digital asset and versioning management), supported by different technologies [1, 5, 8-21]. Existing e-commerce toolkits provide the system administrator/moderator of different capabilities of content management (form, formalize, organize, add, edit, delete). But they do not solve the automatic processing problem of information resources [1-7]. Therefore, for the content lifecycle is a necessary tool that implements process of automatic formation, management and support of commercial content [20].

Content is characterized by a updatable/modifications time and has set specific properties. Content volume is measured in units of information amounts. Content quantity/quality characterizes the degree of user interest to information resources, where it is [1, 4]. Economic content is part of the economic activity of the e-business subject. Content market Internet-based information technologies of knowledge management are tools that facilitate the e-business operation in the commercial content distribution and its profitability increase to e-commerce subject [1]. Commercial content is the object of sales/purchase between the e-commerce participants [1]. For example, information blocks are divided into syndicates (exchange rates, weather forecast unit), announcement of the other sections/resources content (with reference), reference information (date holiday, events announcements, train schedules), entertainment services and information (the day joke), advertising, information partners buttons and links, statistics buttons.

## RECENT RESEARCH AND PUBLICATIONS ANALYSIS

The model authors of the commercial content lifecycle offer and describe the several stages of the properties set that are supported by various technologies and processes. In some models of content lifecycle is project/content/resource management concept, information architecture, content strategy, semantic print. Different authors suggest different stages of the content lifecycle [8-21]. The main stages (content creation, development, revision, distribution and archiving) are present in almost all of the proposed models [8-21]. The processes, activities, status and role lifecycle of content management are different in models depending on organizational strategies, needs, requirements and capabilities of this models [1, 20]. The examined models of the content lifecycle not solve the problem of its formation and support. They also decide not all of management problems: the content set representation to the end user according to his request, history or information portfolio; automatic creation digests and information portrait; thematic subjects and content duplication identification; interconnection table's construction and content ratings calculation; data collection from various sources and its formatting; content keywords identification; content categorization and selective distribution.

Existing e-commerce systems do not support the entire lifecycle of commercial content. They also do not solve the basic problem of information resources processing as the content formation and support (Table 1).

**Table 1. Lifecycle classification of the commercial Web content in electronic commerce systems**

| Content formation | Content management | Content support |
|---|---|---|
| Content collection/creation (information collecting from various information resources and its formatting). | Content editing (a database formation, their rotation and access organization). | Content structurization (content themes identification, concepts relationship building). |
| Content systematization (keywords identification, content duplication, categorization, digests formation). | Content processing (personalization and statistical analysis of users work). | Content analysis (information portraits formation of user and target audience). |
| Content distribution (moderators ranking and selective distribution of content). | Content presentation (search in the database, output forms generation, interaction with other databases). | Content generalization (ratings calculation, new events identification, their monitoring and clustering). |

The electronic content commerce system receives complete information about visitors of information resource that allows a marketing system building. The electronic content commerce system allows to collect maximum statistical information for analysis and to use it efficiently (possible to identify information resource location for optimal placement of advertising information, to automate the progress of the advertising campaign).

## OBTAINED SCIENTIFIC RESULTS ANALYSIS

The process of commercial content formation for an information resource provides a link between a input data set from different sources and formed commercial content set, that stored in the corresponding database of the electronic content

commerce system, namely $Sourse(x_i) \rightarrow x_i \rightarrow X \rightarrow Formation(u_f, x_i, t_p) \rightarrow c_r \rightarrow C \rightarrow DataBase(C)$, where $Sourse(x_i)$ – the content $x_i$ source, $x_i$ – $i$-th content from the source, $X$ – the data set from relevant sources, $Formation(u_f, x_i, t_p)$ – operator of commercial content formation from the $x_i$ content of source to the fixed time $t_p$ with the conditions $u_f$ of content formation, $c_r$ – formed under conditions $u_f$ of $r$- th commercial content, $C$ – set of formed commercial content, $DataBase(C)$ – database of content content. The content sources types for subsystems of commercial content formation are: address list of information resources with confidence and the necessary data, address list of information resources with a subscription to their content, a content set from moderators and content authors, a queries list with keywords for search engines. Content formation subsystem provides information collection from various information resources and its formation, keywords identification and duplication finding, digest formation, content categorization and selective distribution. The main stages of the commercial content formation in electronic content commerce systems are a content formatting, categorization and distribution. They have such scheme of connections: *a content source $\rightarrow$ the content collection/creation $\rightarrow$ database $\rightarrow$ $\rightarrow$ content formatting $\rightarrow$ database $\rightarrow$ keywords and concepts identification $\rightarrow$ database $\rightarrow$ content categorization $\rightarrow$ database $\rightarrow$ duplicate content detection $\rightarrow$ database $\rightarrow$ content digest formation $\rightarrow$ database $\rightarrow$ selective distribution of content $\rightarrow$ moderator*. A model of the content formation in electronic commerce content systems is presented as

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

where $X = \{x_1, x_2 \ldots, x_{n_X}\}$ – the set of input data $x_i \in X$ from different information resources or the moderators at $i = \overline{1, n_X}$ ; $Gathering$ – operator of content collection/creation from various sources; $Formatting$ – content formatting operator; $KeyWords$ – the operator of content keywords and concepts detection; $Categorization$ – automatic categorization operator of content; $Backup$ – the operator of duplicate content detection; $BuDigest$ – the operator of content digest formation; $Dissemination$ – selective distribution operator of Web content; $T = \{t_1, t_2 \ldots, t_{n_T}\}$ – the transaction time $t_p \in T$ of the content formation at $p = \overline{1, n_T}$ ; $C = \{c_1, c_2 \ldots, c_{n_C}\}$ – Web content set $c_r \in C$ at $r = \overline{1, n_C}$ .

The commercial Web content formation describes the operator as $c_r = Formation(u_f, x_i, t_p)$, where $u_f$ – the set of conditions for content formation, ie $u_f = \{u_1(x_i), \ldots, u_{n_U}(x_i)\}$. Commercial Web content is presented as follows

$$c_r = \left\{ \bigcup_f u_f \Big| (x_i \in X) \wedge (\exists u_f \in U), U = U_{x_i} \vee U_{\overline{x_i}}, i = \overline{1, m}, f = \overline{1, n} \right\}.$$ It passes following stages of the transformation wiht a

set data in set of relevant, formatted, classified and validated content: $x_i \in X \rightarrow Gathering(u_f, x_i, t_p) \rightarrow Backup(c_r, u_b, t_p) \rightarrow Formatting(c_r, t_p) \rightarrow KeyWords(c_r, t_p) \rightarrow Categorization(c_r, t_p) \rightarrow BuDigest(c_r, t_p) \rightarrow Dissemination(c_r, t_p) \rightarrow c_r \in C$ . Lexical-grammatical and semantic-pragmatic construction analysis of the text used in the automatically categorization module of content. Its main task is to find textual information in a content stream using content analysis. This information should answer most of the site theme and user needs. Steps performing listed in Table 2 leads to the thematically matched sets formation of textual information. They accumulate information on all aspects coverage of the researched problem, given the opinions and views diversity. The main form of information presentation and preservation is a natural language. Therefore, the effectiveness of the content categorization depends on the automatic processing problem of textual content. Its ultimate goal is their content recognition (Table 3).

**Table 2. The main stages of algorithm in the content categorization module**

| Name | Stage assignment |
|---|---|
| Preparation | Analysis topic, purpose and object determination, its chronological and geographical scope, selection principles. |
| Classification | Classifier formation of key quotes and instructions selection for encoder. |
| Coding | Fragments encoding of textual information. |
| Archiving | Fragments saving of text information in the database. |
| Analysis | Automatic processing of text information fragments. |

**Table 3. The main stages of text processing for content recognition**

| Processing | Stage assignment |
|---|---|
| Formal | It is the text fragments transformation without reference to its content analysis. This step is the foundation of all existing information technology in existing systems of content processing. Morphological information provides access to content, mediated by the ratio of content units with expression units. |
| Thematic (semantic) | At this stage is identified the individual elements content and logical-semantic relations among them for the message semantics presentation. It is the additional semantic information use (which is not expressed explicitly in the text). This stage is a field for theoretical and experimental studies of automatic semantic analysis. |
| Syntactic | Automatically to every word syntactically are tied word forms in a sentence if there the lexical grammatical and grammatical information. |
| Morphemic | At this stage carried morphemic segmentation of text. This prefixes allocation are possible without the language parts knowledge, and suffixes – no: require different sets and procedures of suffix clipping for nouns, verbs, adjectives, adverbs. |

A mandatory integral part in linguistic software of content processing is morphological analysis. So this is the space definition for information text units in the morphological system of the appropriate language and a word forms identition for a single lexeme (Fig. 1). As a result of the morphological analysis of each word-form of content is credited the language parts codes and grammatical categories value (genus, number, case, type, time, person, etc.). This information nature, its scope and methods (which is stablished morphological information) depends on the study purpose (within which provides analyzes) and orientation to the nature of the analyzed texts. The analysis is present at all stages of text analysis because morphemic, syntactic and semantic analysis can not do without the language parts identification. In Fig. 2 presents the automatic coding implementation of text words. It is codes attributing of grammatical classes to them associated with grammatical classification. The analysis object is the structure of words, inflection forms, expression methods of grammatical values.
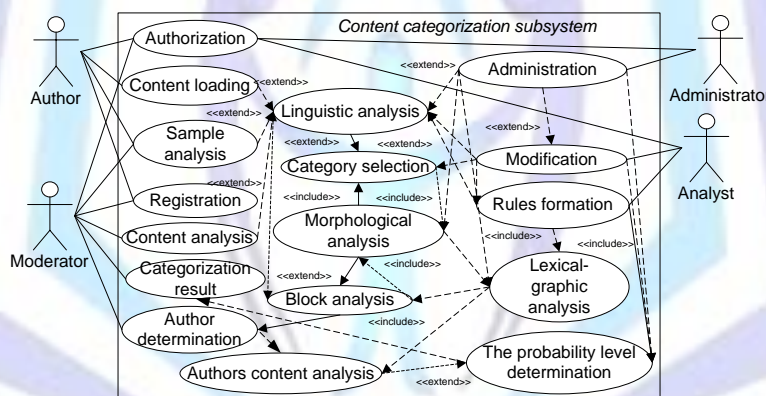


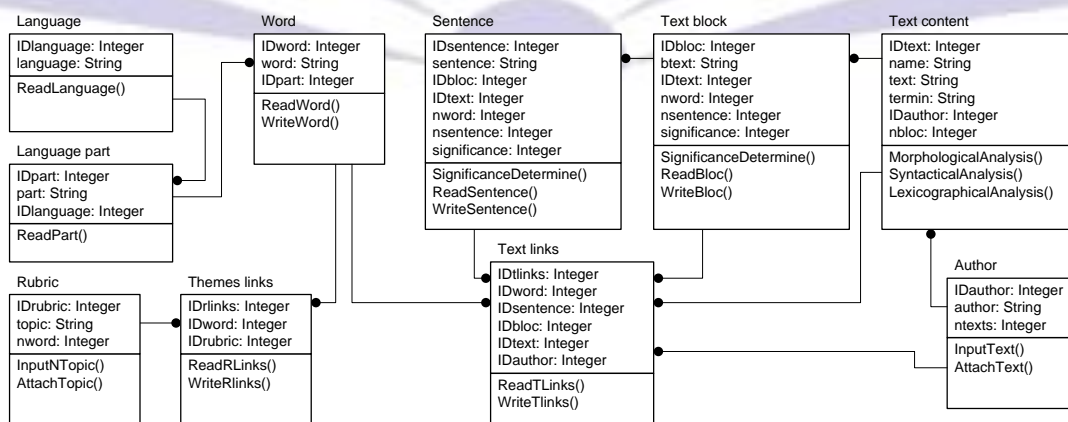**Fig 1: Use Case Diagram for the process of information product categorization**



**Fig 2: Class diagram for the process of information product categorization**

Automatic content analysis is divided into three parts: automatic selection of the base word forms in the text; basics search in the basics dictionary; the word structure comparison with data on its basis contained in the basics dictionary. Each word form in text analyzes using pre-concluded dictionaries of basics, roots, prefixes, suffixes, inflections. In an analysis is the

words meanings and syntagmatic relations identification between words of input text. Analysis tools are the basics dictionary, word endings dictionary, homonymous basics dictionary, semantic and syntactic compatibility table of prepositional constructions components, lexical homonyms removal, semantic analysis of non-prepositional constructions in nouns, semantic compatibility table of nouns and adjectives, automatic analysis algorithms that determine a sequence of inspections and appeals to the dictionary and tables. Procedure for automatic indexing is divided into blocks. It works consistently: morphological analysis, syntactic parsing, semantic and syntactic parsing of prepositional structures and semantic query record variation (Fig. 3).
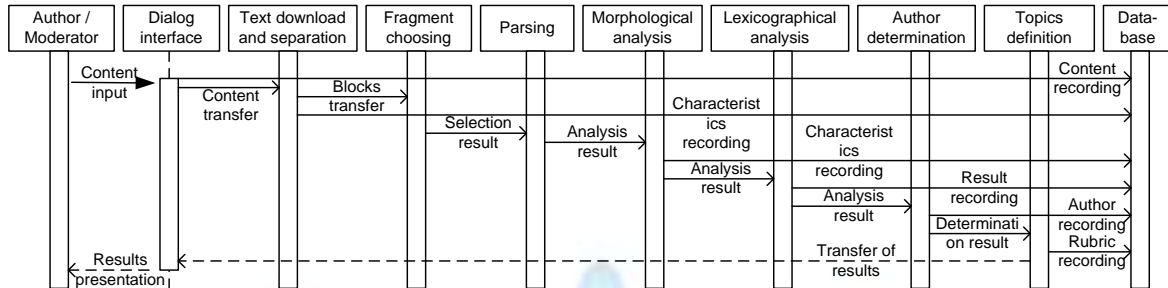


**Fig 3: Sequence diagram for the process of information product categorization**

The system is based on dictionaries and means for their service: system of the words text division on inflection and basics; equivalences dictionary (thesaurus), which is designed to replace the equivalent words with one or more numbers of concepts (they serve as content identifiers instead of words basics); thesaurus as a concepts hierarchy, which provides a search for notion of the general or associated concept with it; dictionaries of statistical and syntactic word combinations; dictionaries service system. Automatic indexing system consists of the following steps: informative words and phrases selection from the text; abbreviations deciphering; words replacing with basic descriptors to the descriptor code; ambiguity removing. Indexing procedure consists of two phases: using descriptive vocabulary; using automatic information retrieval thesaurus (Fig. 4 a).
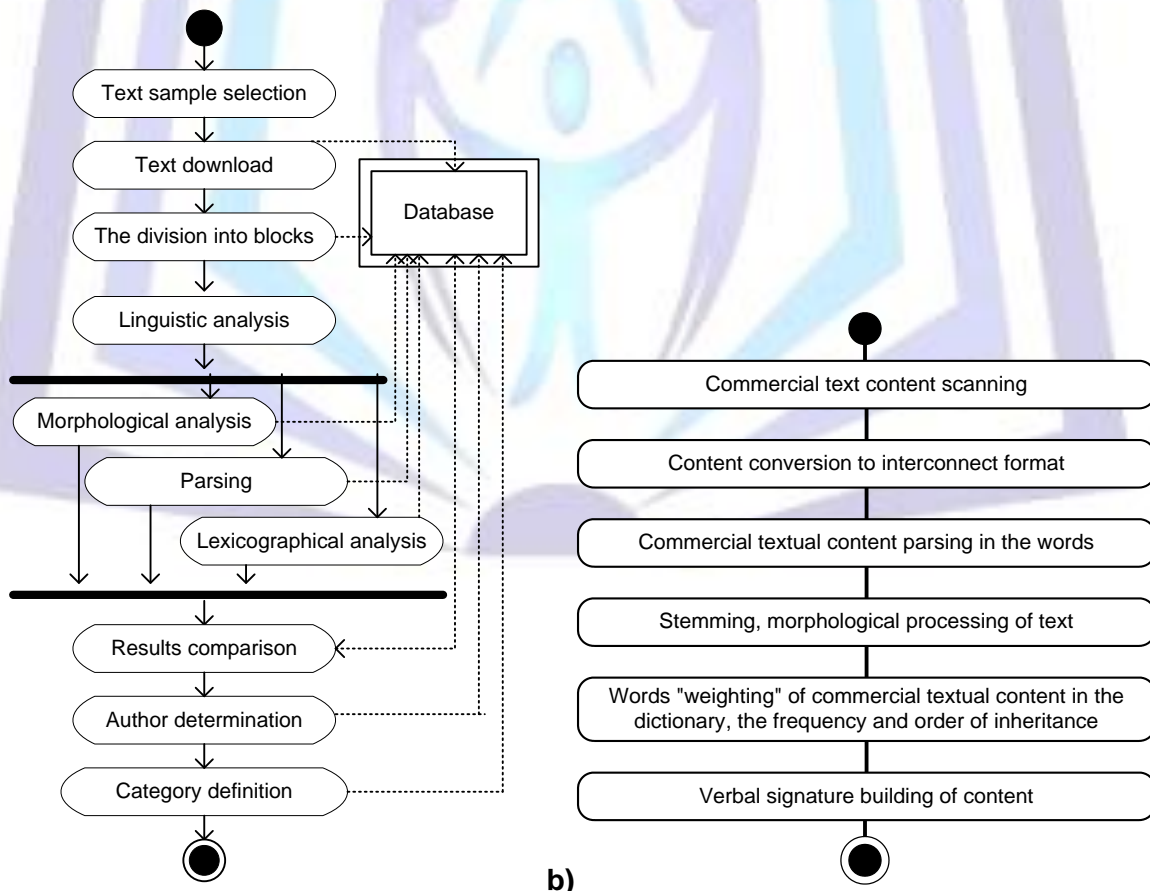


**Fig 4: a) Activity diagram for the categorization subsystem and b) duplicates detecting of information product**

Descriptors dictionary has the table structure. It consists of three columns: words basics, descriptors sets that are assigned to each basis; grammatical features for descriptors. Content categorization is implemented with partial morphological analysis.

1. It is the hierarchical structure construction for properties of each lexical unit in language (it shoulds contain grammatical and semantic information) and the lexicon designing with a hierarchical organization of properties types (where each type-descendant inherits and overrides ancestor property).

2. Unification – base mechanism for syntactic structures constructing.

The principle of keywords detecting in content (terms) is based on Zipf's law. It is reduced to the words choice with an average frequency of use (the most use words are ignored by the "stop-dictionary" and rare words from text information is not included). Duplicate content removal (Fig. 4, b) in the content flow is not always necessary. There are several tasks that use the duplicate content fact from different sources. For example, it is for the important content determining (if the message is duplicated many on sites) or for the PR-campaigns effectiveness determining (press releases republications counting, etc.). In the model constructing investigated the level of content duplication from single Web-sites that have links to other sites source. The content duplicate detecting in textual information at ECCS perform based on linguistic and statistical methods. They are to identify common terms in different content, which form a chain of verbal content signature. The module of a selective content distribution implements load balancing between the moderators in ECCS. These systems are being actively developed, expande their area of influence, encompasse a growing audience of readers and multiply the number of operational information. Typically, the content distribution is performed by moderators. Automatic selective content distribution reduces the time and resources to further the ECCS functioning.

Content contains a theme and a short description of the event (digest). Selective content distribution module automatically distributes digests between moderators according automatically ranking of their work quality.

In Fig. 5 is submitted class diagram of the automatic distribution of digests between moderators. This diagram has its main classes, data and methods, and the relationship between them.
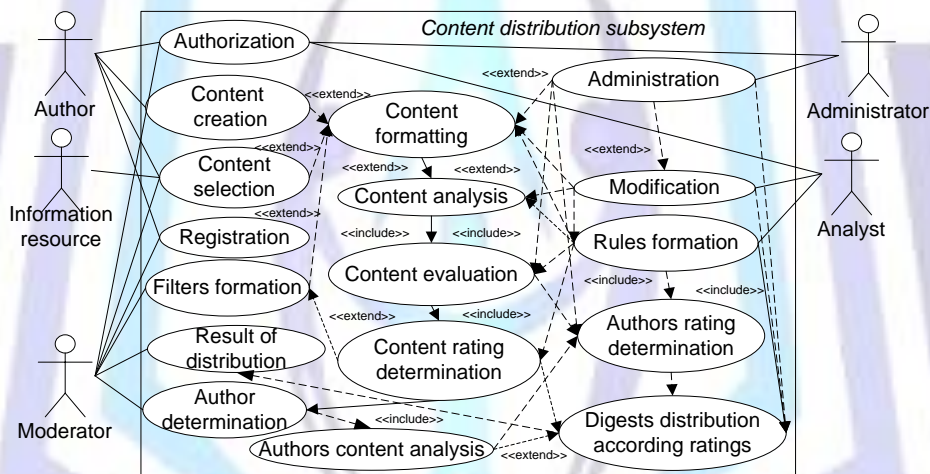


**Fig 5: Use case diagram for the process of information products distribution**

In Fig. 6 is posted cooperation diagram. It illustrates the moderator interaction from the modules of material evaluation, journalists ranking and digests distribution.
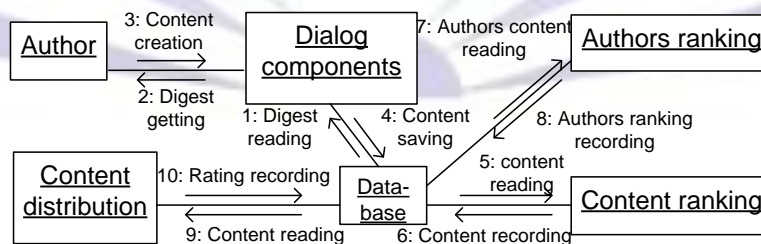


**Fig 6: Cooperation Diagram for the process of information products distribution**

Ranking module evaluates the quality of the work on the criteria set. This ensures quality work objectivity and stimulating. The moderator fate is reduced in such routine work as the same type of information distribution, its sorting, evaluation and analysis. This frees up human resources from mechanical work and improves the results of the objective assessment of the executed tasks quality. At first module gets ready digests with remote network resources by using RSS. Further digests are distributed among the moderators by their ratings. The workers with the highest rating receive the first opportunity to choose digests for processing. The employees rating are determined dynamically. It points to the work productivity and efficiency of each employee separately. It is influenced by criteria such as the unique content percentage, the content views number, user score and residence time on the page with content. This allows objectively assess the moderators work quality. In sequence diagrams (Fig. 7) are presented the objects interaction, sorted by time of their manifestation. The diagram gives a clear understanding of the processes occurring in the system.
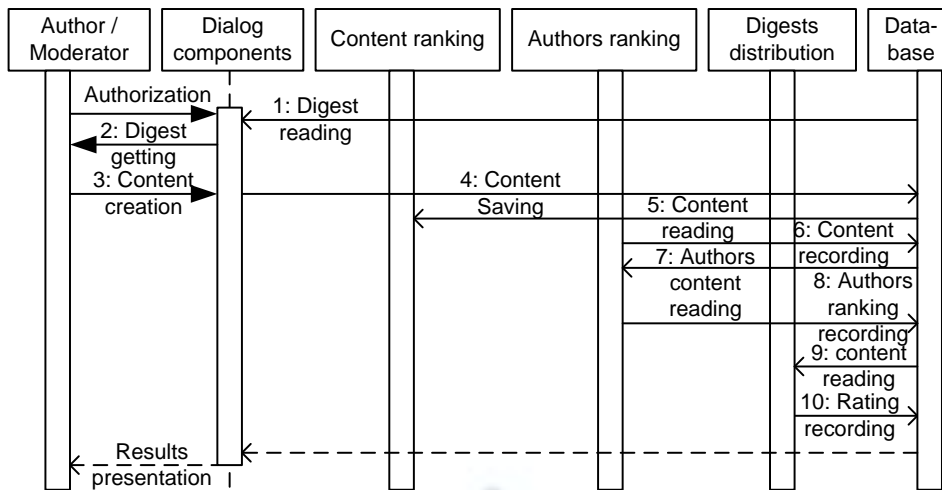
**Fig 7: Sequence diagram for the process of information products distribution**

In Fig. 8 is presented class diagram for the module of digests distribution, which reflects the main classes of the system.
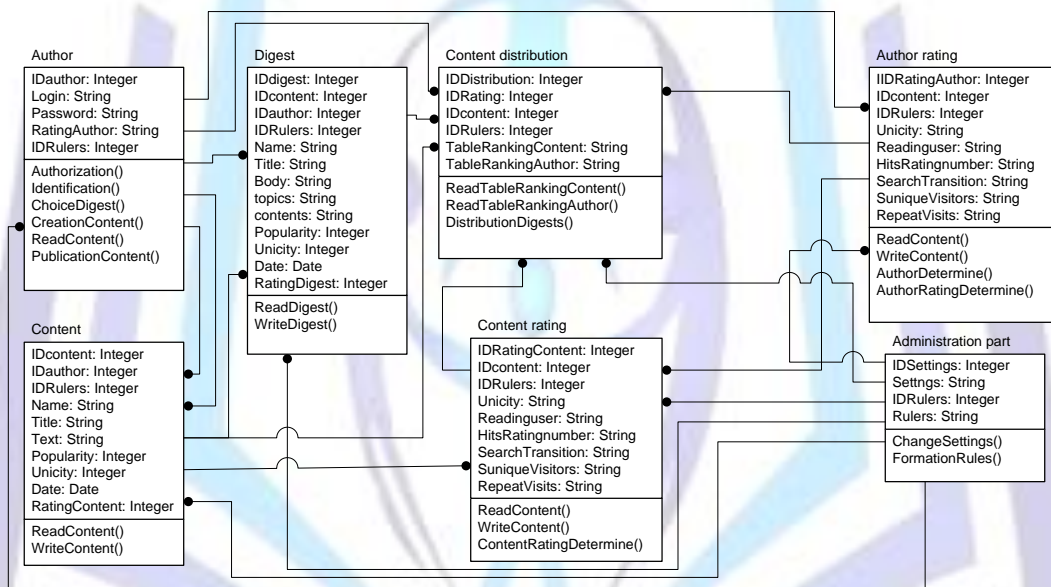


**Fig 8: Class diagram for the process of information products distribution**

The content contains theme and digest. Content distribution subsystem selectively sends digests between authors by the quality ranking of their work. Ranking subsystem assesses the work quality on criteria set. This ensures quality work objectivity and stimulating. The moderator work fate as a similar data distribution, their sorting, evaluation and analysis is reduced. This reduces the resources using, reduces the content creation and improves content characteristics due to objective quality evaluation of executed tasks. With the percentage of text uniqueness evaluate the author work quality of this content and records the obtained evaluation in the rankings table. The content volume increase leads to a more accurate assessment of the quality and performance of each content author's. The number increasing of evaluation criteria allows to cover wider range of the author/moderator work aspects. In Fig. 9.a is presented module activity diagram. That diagram shows the specification of executable behavior in the form of coordinated series and parallel execution of subordinate elements. The values in Fig. 9.b show the hierarchy elements priorities in terms of the goal. The greatest weight given to the «Unicity» criterion that objectively reflects the moderator work quality. The next the weight is «Reading time», which reflects the users interest extent in the content. «User evaluation» and «Visiting» criteria carry less weight because they do not have the trust due to the figures falsification ease.
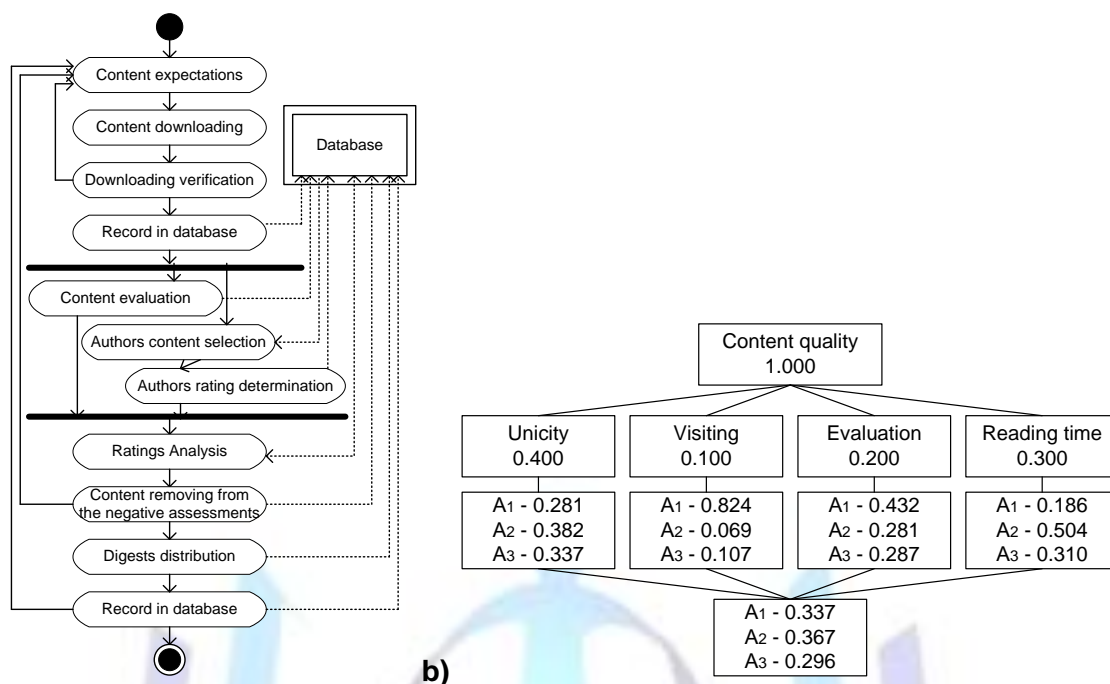
Fig 9: a) Activity diagram for the process of the information products distribution and b) the criteria hierarchy for moderators rating evaluation

The main results are implemented in the Internet projects "Vgolos" (Lviv, vgolos.com.ua), "Victana" (Lviv, victana.lviv.ua), "Tatjana" (Kherson, tatjana.in.ua), "Presstime" (Lviv, presstime.com.ua), "AutoChip" (Vinnitsa, autochip.vn.ua), "Fotoghalereja-vysocjkykh" (Lviv, fotoghalereja-vysocjkykh.com), "Kursyvalyut" (Lviv, kursyvalyut.com), "Dobryjranok "(Lviv, dobryjranok.com). In Fig. 10 are presented the comparative characteristics of the work results of the developed systems for the period 10-11.2012. They are obtained from Google Analytics (accounts.google.com), which provides advanced features of data analysis. Google Analytics is a free service for statistics analysis of Web-sites visitor. It allows to evaluate Web-site traffic and marketing measures effectiveness, such as for the newspaper aloud (Fig. 10).
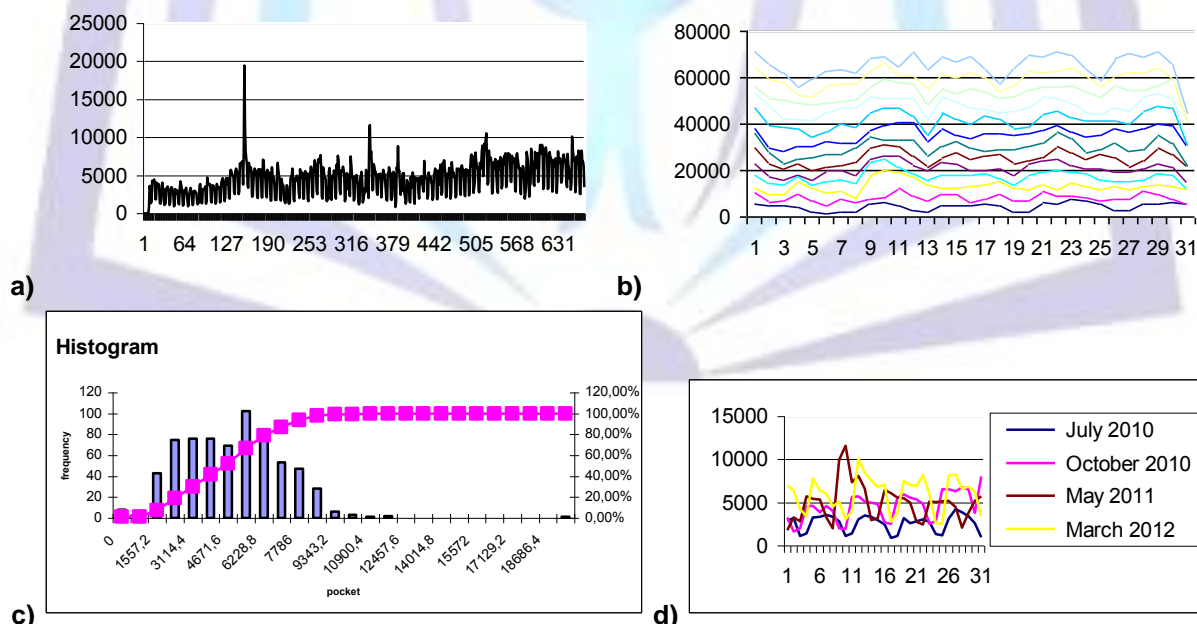


Fig 10: Monthly visitors distribution in years 2010-2012

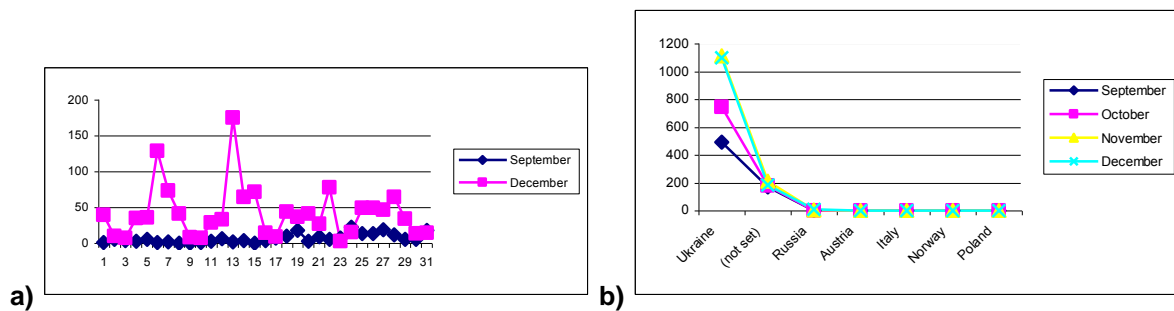In Fig. 11-12 are presented the work results of the developed system «Good morning, accountant!» (http://dobryjranok.com/).

a)                                                                b)

**Fig 11: Distribution of a) the resource visiting and b) visiting from various countries in the world**



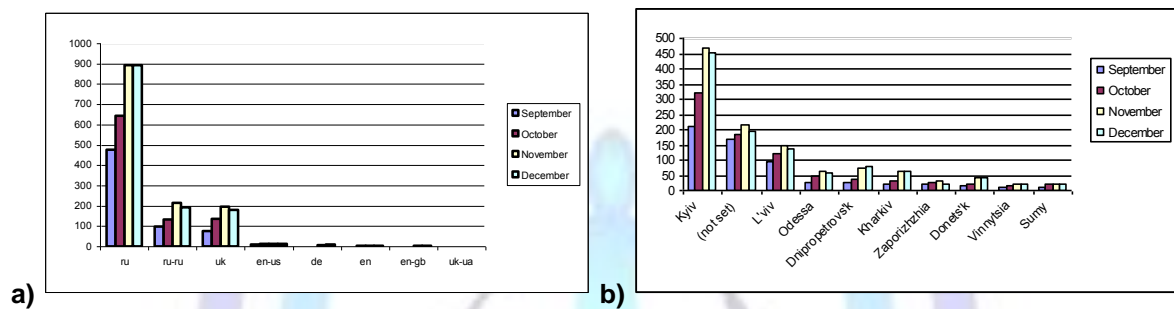a)                                                                b)

**Fig 12: Distribution of a) traffic speech and b) the resource visiting from different cities**

The results are obtained from Google Analytics as graphs and charts. They imply that the all stages presence of the content lifecycle in the resource significantly increases the amount of visits and unique users. The excessive results causes of the standard content search are divided into two different categories: duplication and inconsistency. It is essential that the content belonging to the doubles has completely objective nature and is determined automatically on the basis of formal criteria.

## CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER SCIENTIFIC STUDIES

In the article are explained terms, concepts, categories, nature, content, scientific and theoretical foundations of the content commerce systems. There are explored new theoretical and practical directions and methodological approaches for e-commerce organization. In the article are analyzed the regularities, trends and phenomena that are characteristic for content-commerce organization. In the work are researched organizational principles, structures, forms and methods of organizational activities of electronic content commerce systems. The authors have substantiated the principles, rules, means of organization, perspective models, and organizational development priorities for similar systems designing. In the article are defined contents of organizational action, strengthening methods of organizational impact to improve of commercial activity efficiency and functions implementing in the electronic content commerce systems. Here are described the basic tools of electronic content commerce. From the perspective of systemic approach are analyzed the principles and developed a model of information resources processing in the electronic content commerce systems. This made it possible to develop methods of commercial content formation, management and distribution. In the work are developed a complex method of the content formation. This makes it possible to create a means of information resources processing and implement the module of automatically formation of Web content. There is established an operational method of content management, making it possible to implement Web content management module. Also in the article is implemented a complex of content distribution method that makes it possible to develop Web content distribution module.

## REFERENCES

[1] Берко А., Висоцька В., Пасічник В., "Системи електронної контент-комерції", Монографія, Л: НУЛП, 2009.

[2] Ермаков А., Киселев С., "Лингвистическая модель для компьютерного анализа тональности публикаций СМИ", Компьютерная лингвистика и интеллектуальные технологии., М: Наука, 2005, 282-285.

[3] Іванов В., "Контент-аналіз: Методологія і методика дослідження ЗМК", Наук. ред. А. Москаленко, Київ, 1994.

[4] Клифтон Б., "Google Analytics: профессиональный анализ посещаемости веб-сайтов", М: И.Д. Вильямс, 2009.

[5] Пасічник В., Висоцька В., Щербина Ю., Шестакевич Т., "Математична лінґвістика. Книга 1. Квантитативна лінґвістика", Львів: "Новий Світ – 2000", 2012.

[6] Сорока М., Танатар Н., "Використання методу контент-аналізу при створенні автоматизова-них інформаційних систем", Бібліотека. Наука. Культура. Інформація: Наукові праці НБУВ, 1998, № 1, 318-322.

[7] Федорчук А., "Контент-мониторинг информационных потоков", Б-ки нац. акад. наук: пробл. функционирования, тенденции развития, Киев, 2005, № 3, http://www.nbuv.gov.ua/articles/2005/05fagmip.html.

[8] Boiko B., Content Management Bible, Hoboken, 2004.

[9] CM Lifecycle Poster, "Content Management Professionals", Retrieved 20 July 2010, http://www.cmprosold.org/resources/poster/.

[10] Doyle B., "Seven Stages of the CM Lifecycle", EcontentMag.com, 2010, http://www.econtentmag.com/Articles/ArticleReader.aspx? ArticleID=13554&AuthorID=155.

[11] Hackos J. "Content Management for Dynamic Web Delivery", Hoboken, NJ: Wiley, 2002.

[12] Halvorson K. "Content Strategy for the Web", Reading, Mass: New Riders Press, 2009.

[13] McGovern G., Norton R., "Content Critical",. McGovern, Upper Saddle River, NJ: FT Press, 2001.

[14] McKeever S. "Understanding Web content management systems: evolution, lifecycle and market", Industrial Management & Data Systems (MCB UP), 2003, 103 (9), 686–692.

[15] Nakano R., "Web content management: a collaborative approach", Boston: Addison Wesley Professional, 2002.

[16] Woods R., "Defining a Model for Content Governance", 2010. www.contentmanager.net/magazine/article_785_defining_ a_ model_for_content_governance.html.

[17] Rockley A., "Managing Enterprise Content: A Unified Content Strategy", Reading, Mass: New Riders Press, 2002.

[18] Stone W. "Plagiarism, Duplicate Publication and Duplicate Submission: They Are All Wrong!", IEEE Antennas and Propagation, Aug. 2003, Vol. 45, № 4.

[19] "The Content Management Possibilities Poster", Metatorial Services, Inc., Retrieved 20 July 2010, http://metatorial.com/pagea.asp?id=poster.

[20] Vysotska V., Chyrun L., "Commercial Web Content Lifecycle Model", CSIT'2011, LPNU, Lviv, 2011, 160-163.

[21] "Web Content Lifecycle", State Government, Victoria, Australia, 2005., http://www.egov.vic.gov.au/victorian-government-resources/reports-victoria/web-content-lifecycle-and-content-management-roles/1-web-content-lifecycle.html.

## Author' biography with Photo

Victoria Vysotska work assistant on Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine. She has obtained Masters degree in Computer Application from Lviv Polytechnic National University. Victoria has over 10 years of teaching, about 70 publications in various national and international journals and conferences. She has a scopes of scientific interest as information systems and networks, ecommerce, information resources, commercial content, content analysis, content monitoring, content search, electronic content commerce systems, software systems, models, algorithms, analysis, methods and strategies of systems design.

Lyubomyr Chyrun work docent on Software Department, Lviv Polytechnic National University, Lviv, Ukraine. PhD is in 2007. He has over 10 years of teaching, research and consultancy experience and has been key note speakers and about 50 publication in various national and international conferences in Ukraine and abroad. He has a scopes of scientific interest as software, information systems and networks, ecommerce, information resources, commercial content, content analysis, content monitoring, content search, electronic content commerce systems, software systems, models, algorithms, analysis, methods and strategies of systems design.