Informational resources processing intellectual systems with textual commercial content linguistic analysis usage constructional means and tools development

L. Chyrun[1], V. Vysotska[2], I. Kozak[3]

[1] Software Department, Lviv Polytechnic National University; e-mail: chyrunlv@mail.ru
[2] Information Systems and Networks Department, Lviv Polytechnic National University, e-mail: victoria.a.vysotska@lpnu.ua
[3] Applied Linguistics Department, Lviv Polytechnic National University; e-mail: ivan.kozak.lp@gmail.com

Received February 18 2016: accepted April 16 2016

Abstract. The article content lies in solving the important applied scientific problem of the informational resources processing intellectual systems (IRPISes) with textual commercial content linguistic analysis usage creation. The IRPISes functioning mathematical ensuring was developed. The IRPISes construction means and methods will be developed on the basis of created mathematical models. Such systems have the widespread usage, in particular for the forming, managing and maintenance of the expanding content volume in Internet, running e-business, during the online and offline content realization systems, cloud storage and cloud computing.

The increase in the content volume causes the proper quality and productivity evaluation of the very content author. The increase in the evaluation criterions allows covering the broader aspect range of any author's / moderator's work.

Key words: informational resource, commercial content, content analysis, content monitoring, content search, e-content commerce system.

INTRODUCTION

The active Internet development promotes the increase in operative industrial/strategic operating data and informational services via up-to-date IT e-commerce realization requests. The commercial content is the documented information prepared according to users' requests. For today the e-commerce is the objective reality and long-range business process. Internet is the business environment and commercial content is the most requested product with sales inside and the main content e-commerce processes object. The commercial content can be ordered, registered, paid and got online as goods strait away. The whole commercial content diversity (viz. scientific and publicistic articles, music, books, films, photos, software etc.) is sold via Internet. The well-known corporations to sell electronic content commerce are Google via Play Market, Apple – Apple Store, Amazon – Amazon.com. The informational resources processing in the informational resources processing intellectual systems (IRPISes) allows to receive the hot and objective data about the system operating and for financial market content segment competitive level evaluation; to estimate the level of competitors, their competitiveness onto the content expansion financial market [4, 6, 13-15, 17].

The majority of decisions and researches were performed on the concrete projects' levels. IRPISes are built on the close basis as one-time projects. Modern IRPISes are oriented on the outside the system commercial content realization. The IRPISes designing, creation, implementation and maintenance are impossible without usage of modern commercial methods and informational technologies to form, manage and support the commercial content. The main categories of the informational resource users/characters (customers, working groups managers and administrators) fix the informational resource design and decision making process. IRPIS has obligatory the informational resource with the content catalogue (with the possibility to search) and requisite interface elements for the registration data entering, the order proceeding, payments execution via Internet, delivery request (e-mail / on-line), receiving information about the company and the on-line help. The whole content managing process is recorded in the content maintenance subsystem to form the IRPIS functional statistics and propositions in content popular subjects list for the content forming subsystem [1-2].

MATERIALS AND METHODS

The informational resources working-out technology development is topical because of the factors such as the theoretical arguments insufficiency of the commercial content torrents processing methods and the IRPISes' informational resources processing software unification necessity. The practical factor of the informational

resources processing in IRPISes is related to the solution of such tasks as forming, managing and maintenance of commercial content growing amount in Internet, e-business fast development, growing possibility of the access to Internet, informational goods and services widening, increase in commercial content request. Principles and informational technologies of e-commerce are used for Internet-shops creation (selling of eBooks, software, video, music, movies, pictures), online systems (newspapers, journals, remote education, publishers) and offline content sales (copywriting services, Marketing Services Shop, RSS Subscription Extension), cloud storage and cloud computing. The major global producers of informational resources proceeding means (such as Apple, Google, Intel, Microsoft, Amazon) work in this area. The theoretical factor of informational resources processing in the IRPIS is connected with the commercial content processing IT development. The electronic informational torrents processing mathematical models were explored and developed in the scientific works of D. Lande, V. Furashev, S. Braichevskyi, O. Hryhoriev. G. Zipf proposed the words frequency distribution empiric rule of natural language in a textual content for its analysis. The content life cycle models are worked out in the works of B. Boiko, S. McKeever, E. Rockley. M. Weber, J. Kaiser, B. Glaser, A. Strauss, H. Lasswell, O. Holsti, V. Ivanov, M. Soroka, A. Fedorchuk introduced and developed the content analysis methodology for textual data arrays processing. The methods of textual information intellectual processing proposed in the works of V. Kornieiev, A. Harieiev, S. Vasiutin, V. Raikh. The EMC, IBM. Microsoft, Alfresco, Open Text, Oracle and SAP corporations has developed the Content Management Interoperability Services specifications for the Web-services interface to provide the e-business content managing systems cooperation. From scientific side of view this IT segment is investigated not enough. The every unique project is realized practically from the very beginning and, in fact, based on one's own ideas and decisions. The essential theoretical grounds, researches, deductions, recommendations, generalizations for the IRPISes designing and informational resources processing in such systems are extremely little regarded in the literature. The necessity in the analysis, generalization, and founding of the e-commerce realization and the IRPIS construction existing approaches has arisen. There is the actual task to create the technological means system based on the theoretical foundation of the IRPISes' informational resources processing methods, models and principles built on the 'open systems' principle which allows to control the process of the commercial content realization increase.

The analysis of the mentioned above factors let make the decision about the existing of some contradiction between the quick IT and IRPISes development and expansion on the one hand and comparatively small amount of scientific works towards this topic and their locality on the other. This contradiction causes the problem of the e-commerce innovational development through proper newest progressive IT creation and establishment restrain, which has the negative influence on this market sector growth. The task of the e-commerce informational resources processing scientifically grounded methods elaboration and the creation of the IRPISes creation, expansion and stable development software based on them is actual in the general problem terms. In this work the research to determine the regularities, specifics and dependences of the informational resources processing in the analogical systems (especially for textual commercial content automation processing) was done.

There are the tendencies of the automatic text processing (ATP) usage (Fig. 1) [20-22]:

- The Turing test: the system can be regarded as intellectual when no difference can be seen while talking with it [20-25, 30-31, 39, 41-44],
- The Internet systems (language naturalness) → the information research [4,6,13-15,17,32,46-48].

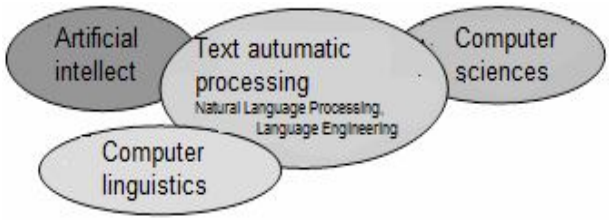


Fig. 1. The automatic text processing

The information research (IR) development fields [20, 27]:

- The fifties – Information Retrieval (IR) [1],
- The nighties – WWW (Google > 8 billion dollars, Yandex > 600 million pages, 2.5 million sites).

The main IR determinations according to [1, 3, 16, 20, 27] (Fig. 2):

- Filing, saving, organization of the information unit and the access to it,
- The focus on the user's informational requests,
- The accentuation on the search of information (not data),

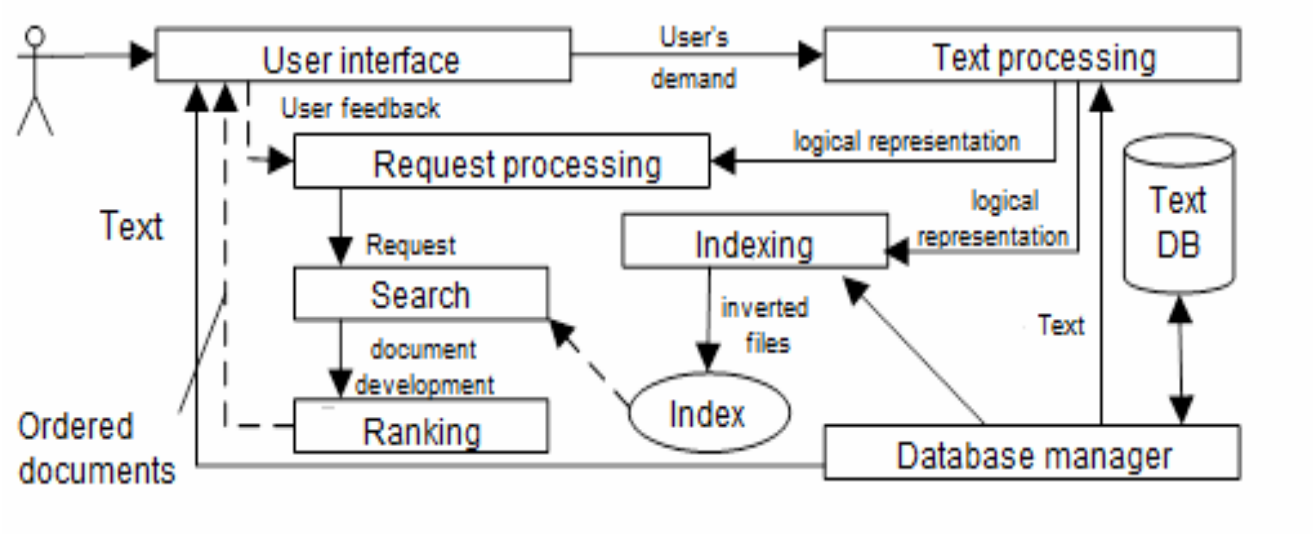


Fig. 2. The IR process

The direct search [20]

S	O	N	I	A	S		D	A	D		I	S		S	M	A	R	T

D	A	D

D	A	D

- Dboyer-Moore with the average complexity O(n/m)

S	O	N	I	A	S		D	A	D		I	S		S	M	A	R	T

D	A	D

D	A	D

D	A	D

D	A	D

- Brute force with the average complexity O(n+m),

The indexation is the process of the document research image creation (logical presentation). Usually it's the inverted index (according to [1,3-4,7-12].

$$\text{Dictionary}\begin{cases}\text{Brutus} & \Rightarrow 2 \to 4 \to 8 \to 16 \to 32 \to 64 \to 128\\ \text{Calpurnia} & \Rightarrow 1 \to 2 \to 3 \to 5 \to 8 \to 13 \to 21\\ \text{Caesar} & \Rightarrow 13 \to 16\end{cases}$$

Postings

The stages of previous ATP [20]:

- The extracting and / or receiving of the text (HTML, PDF…),
- The coding and language determination,
- The fragmentation of words and sentences (tokenization),
- The stop-words abolition,
- The lemmatization (stemming) – bringing the word into the lexicographic form.

Tokenization (the example) [20]:

1. Dates, numbers: 13/03/2014, 1415…;
2. Adverbs: finally, usually, thence, then, e.g.…;
3. The opening speech: in other words, to summarize, apropos…;
4. Prepositions: on the eve of, despite of…;
5. Particles: and yet, as if, like, besides, seems like…;
6. Verbose tokens: Ulan-Ude, New York, Ivan Ivanovych… (collocations);
7. The sentences' scopes: The last winter I.I. Ivanov came to Lviv.

The stop word determination [20]:

- Text = unstructured set of meaningful words ("bag of words"),
- The stop-words are the ancillary word class (prepositions, conjunctions, particles…): ah, aha, wow, yeah, hurrah, aside, along, besides, so, except, instead of, inside, outside, near…

The IR model [20]:

- The documents file method,
- The information request assigning method,
- The propinquity between request and document computation method.

The Boolean IR model [20]:

- The document = the set of words (terms),
- The request = the Boolean expression:

(cat OR dog) and food
Bird ANDNOT soldier

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
Mercy	1	0	1	1	1	1
Worser	1	0	1	1	1	0

The Boolean IR model peculiarities [20]:

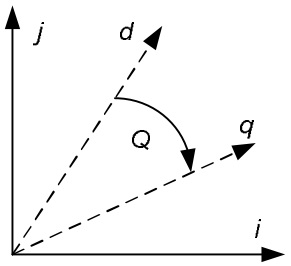
Advantages	Disadvantages
Simplicity	Too 'contrasting' (the document presentation as well as relevance)
Is easy for those who know the logical operators	

The IR vector model [10-12, 18, 20]:

- The document and the request are vectors in the space of words (terms); the importance of the word for the document / request is the vector component,
- The propinquity measure is the cosine of the angle between vectors ($\rightarrow$ ranking)

$$sim(\bar{d},\bar{q}) = \frac{\sum d_i \cdot q_i}{|\bar{d}||\bar{q}|},$$

where: d_i is the i term value in the document, q_i is the i term value in the request [1, 16]:



The term importance is calculated with taking into consideration the next factors [20, 26, 45, 49]:

1. How often does it occur in the document?
2. How often does it occur in the collection?

The $TF \cdot IDF$ approach, where TF - term frequency, IDF - inverse document frequency, so $TF \cdot IDF$ is the basic variant [1, 16].

$$tf_{ij} = \frac{f_{ij}}{\max\limits_k f_{kj}},\ idf_i = \log\frac{N}{n_i},\ w_{ij} = tf_{ij} \cdot idf_i$$

$TF \cdot IDF$, Okapi [20-22].

$$TFIDF_D(l) = \beta + (1-\beta) \cdot tf_D(l) \cdot idf_D(l).$$

- The request processing = the operations with the sets corresponding to words (terms).

The Boolean model example (according to [10-12, 20]):

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
Mercy	1	0	1	1	1	1
Worser	1	0	1	1	1	0

$$tf_D(l) \frac{freq_D(l)}{freq_D(l) + 0.5 + 1.5 \cdot \frac{dl_D}{avg\_dl}},$$

$$idf(l) = \frac{\log\left(\frac{|c| + 0.5}{df(l)}\right)}{\log(|c| + 1)},$$

where: avg_dl is the average document length, c is the collection size, $\beta = 0...1$

The vector model peculiarities

Advantages	Disadvantages
Works nice in the "pure" static collections	Can be easily attacked (spam)
Partial coincidences are conceded	Has low efficiency with short texts

Web [4, 6, 13-15, 17]

- Uncontrolled collection,
- Huge amounts,
- Different formats,
- Diversity (language, topics…),
- Competition (spam),
- Clicks,
- Links! (PageRank)

The base for research quality evaluation is the relevance concept (relevance to the information request), viz. the precision $p = a/b$, the recall $r = a/c$ and the F-limit $F = (p+r)/2pr$, where a is the number of relevant results in the reply, b is the number of all results in the reply, c is the number of all relevant results [20, 27].

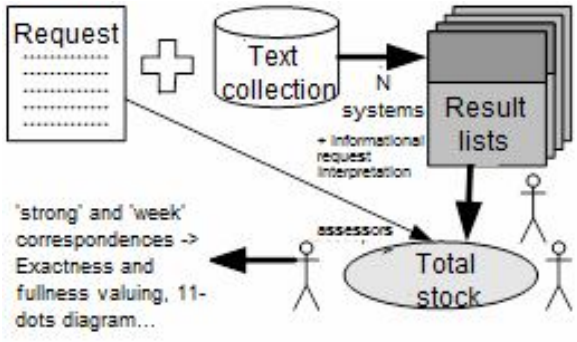


Fig. 3. The 'total boiler' method

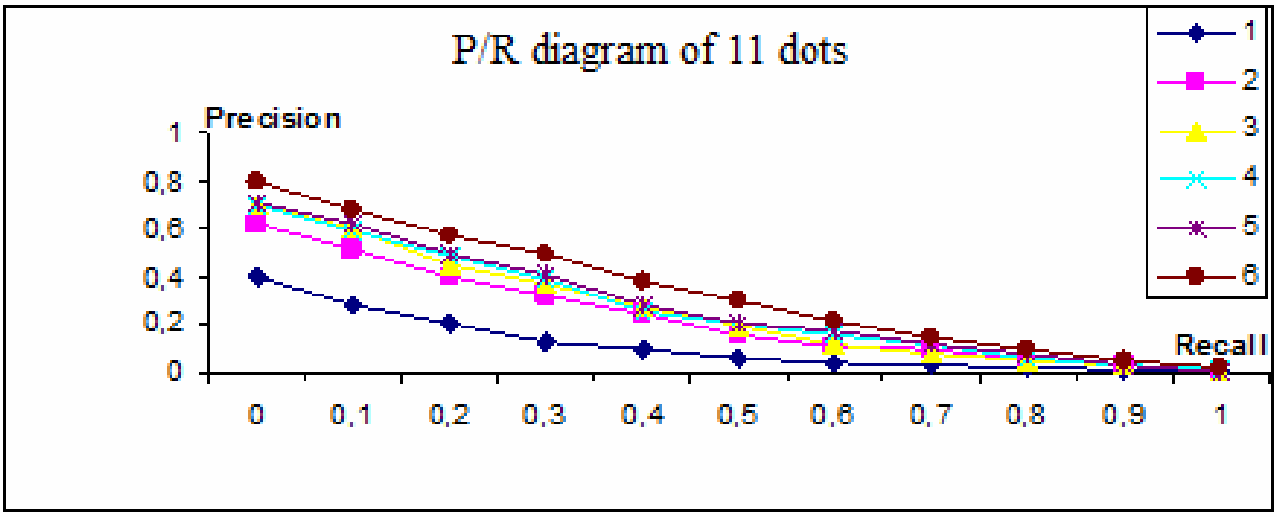


Fig. 4. P/R diagram of 11 dots

RESULTS AND DISCUSSION

The automatic morphological analysis necessity [20, 33-34, 36-38, 50-60]:

- The key words equivalent classes while searching: cat, cats, with cat…,
- The further processing (syntactic analysis, semantic analysis…).

The analysis types [20]:

- Stemming is the stem severance: *wood, wooden, woody, wooded* → *'wood'*, or *system, systematical, systematically, systemize* → *'system'*,
- Unification to the lexicographic form: *dancing* → *to dance, damaged* → *to damage, woods* → *a wood,*
- POS-tagging (part-of-speech): *leaves <N> dancing <V> in <PREP> the <ARTICLE> air <N>,*
- Absolute morphological information: *leaves <N, plural, not-being> dancing <V, present participle> in <PREP> the <ARTICLE, definite> air <N, single, not-being>.*

Stemming is the process of shortening the word to the stem by throwing affixes (like ending and / or suffix) away. The stemming results are similar to the word root finding but still the stemming result may differ from the morphological root of a word. Stemming is used in the morphological analysis and IR. The majority of search systems use the stemming for the slur process, i.e. merging of words into sets (synonyms), which have the similar forms after the stemming. After the stemming process is done the words *active, actively, activity* are led to the form *'activ'*. And the words *loud, loudly, aloud* are processed to the very root *'loud'*. The most active researches in this area were performed by Martin Porter. He had developed the algorithm which became widely popular and in fact became the standard stemming algorithm for the English language only [20].

Search in the table, where all the possible words variants after the stemming are collected [20]. The advantages are the simplicity, quickness and easement of the linguistic exceptions working out. As the disadvantages there may be regarded the search table should contain all the word forms, and, as the result, the algorithm will not work with new words and the table volume may be quite huge. For the languages with the simple morphology e.g. English the research table sizes are quite small, while for Turkish or Ukrainian there exist a huge amount of words with the one root. The table fragment of the *word* → *stemming* model for example for the word *гарний* looks like { *гарна*, *гарне*, *гарний*, *гарним*, *гарними*, *гарних*, *гарні*, *гарній*, *гарнім*, *гарного*, *гарної*, *гарному*, *гарною*, *гарну*} → *гарн*.

The endings and suffixes throwing away is based on the word shortening rules [20], e.g. if the word ends with '*льна*', '*ьна*' is cut off, if it does with '*льне*', '*ьне*' is cut off, if it does with '*льний*', '*ьний*' is cut off, if it does with '*льним*', '*ьним*' is cut off. Those stemming rules number is much smaller than the table with all word forms, thus the algorithm is quite compact and productive. The mentioned rules process correctly the following adjectives with the *word* → *stemming* model, e.g. '*вільна*'→'*віл*', '*мильне*'→'*мил*', '*сильний*'→'*сил*' та '*суспільний*'→ '*суспіл*'. The algorithm may make the mistakes. For example in the word '*пальне*' the wrong form '*пал*' may be instead of *'пальн'*. Taking into consideration the language special features the endings and suffixes throwing away rules set is complex especially for the Slavic languages. As the disadvantages there may be regarded the exceptions (the words which have the variable form) processing. For example words '*криком*' and *'кричу'* should have the *'крик'* form after the stemming process. The algorithm should take this into consideration which causes the rules complications and thus have a negative effect on the productivity.

Lemmatization is based on the word stem finding by the POS tagging (word classes' determination in a sentence) [20]. Then the stemming rules are applied to the word according to its word class belonging. Thus words '*пальне*' (noun) and '*сильне*' (adjective) should be processed with different rules sets. These algorithms offer the high quality and have the minimum mistakes percent if the part of speech discerning rules are correctly listed.

Scholastic algorithms are based on the word stem determination probability [20]. They have the ability *to learn*. The set of logical rules and the search tables are the knowledge base for these algorithms. After the word was processed with the scholastic algorithm there may be several word stem variants within which the algorithm takes the most probable one. For example there is the only one logical rule of the last letters cutting off. The knowledge base is built in the *word→ stemming→ ending* model, e.g. { *популярність → популярн → ість*, *хвилини → хвилин → и*, *добрими → добр → ими* }. The *ending* domain contains the results of the algorithm training based on the knowledge base. For the illustration let's perform the stemming of the word *'львівяни'* using the *word → ends with? → result→ numerical result* model i.e. {*львівяни → ість → по→* 0, *львівяни→и→yes→*1, *львівяни→ими→по→*0}. The only result was got, so the word after the stemming looks like '*львівян*'. But if the word *'відомими'* is processed with this algorithm the result with the *word → ends with? → result→ numerical result* model is ambiguous, i.e. { *відомими → ість→ по→* 0, *відомими → и→ yes→* 1, *відомими → yes→* 1}. The rule complication solves the contradiction because the stemming (within which the a word is shortened more or less) is more preferred.

The hybrid stemming approach combines all the motioned above algorithms. For example the algorithm uses the endings and suffixes throwing away method but on the first stage performs the search in the table. In spite of the search in the table this one contains not all the word forms but only the rule exceptions which are worked out incorrectly by the algorithm that cut off endings [20].

The prefixes cutting off process exists simultaneously with the word's suffix and ending cutting off. Not all prefixes may be disjoined from a word, e.g. the word *'незалежний' ('independent')* will be transformed to the *'залежн' ('dependent')* form, which is the exact antonym. But there are quite a lot of words in which a prefix don't change the main word meaning, e.g. *'проголошую, наголошувати, виголошував'* may be easily shorten to the *'голошу' ('tell')*.

For **the correspondence search** the knowledge base which only contains words' stems is used. In the other words this knowledge base consists of the forms created after the stemming process has been done to the usual words. If we parallel with the search in the table, it's the second column words. The main object of these algorithms is to find the most appropriate word's form in the knowledge base using the inside rules system. The similar length of the word and its stem system may be one of these rules. For example the knowledge base contains two stems '*чорн*' and '*чорняв*'. In comparison with the word '*чорнява*' the first variant has 4 joint symbols ('*чорн*') and the second one does 6 symbols ('*чорняв*'). Thus the algorithm takes the longer variant.

The stemming process within different languages. The first academic works were dedicated only to the English language, but now there are quite a lot of stemming realizations for various languages. The stemming algorithm writing process complexity depends on the peculiarities of the language. For English the stemming is quite a trivial task but for Arabic or Hebrew the task is ten times more difficult. There exist the stemming variants for the Ukrainian language and they are used in the commercial search systems. For this moment the realization of such algorithms isn't free.

Mistakes in the stemming process. Within the stemming algorithms the two types of mistakes are common.

- *Overstemming* – the stemming process which causes the shortening of two different words to the only stem (but this should never happen),
- *Understemming* – the mistake of the contrary matter. Within it words get different stems but they should get the same one.

The stemming algorithms aims to minimize the similar mistakes, but the minimization of one type mistakes may cause an increase of the other ones.

The IRPIS's information source is the set of data with the certain attributes (table 1) which are the objective of the IT operations within their transmutation into content process [2, 5-6, 19-22, 25-29, 32-40, 50-60].

Table 1. The IRPIS's information resources main properties

Name	Property
Heterogeneity	Presence of components with various nature, contents and file format
Coordination	Absence of contradictory and opposite content data
Format accessibility	Accessibility for all users based on the standardized means, methods and interfaces.
Openness	Possibility of the cooperation, data exchange and common usage with any external resources.
Dynamic	Quick actualization according to the system / environment terms
Scalability	Possibility to change the logical / real content volume (values, conceptions and their symbols)
Controllability	Content change / usage identification and the influence of content on the information system (IS) processes.

A result of the one IT usage may be an information resource for the other one. The IT content is the formalized statements and knowledge contained in the IS environment (in spite of the data which exists without any detailed specification of their properties, formalization and normalization methods. One of the significant issues of the IRPIS construction and functioning is the transformation of data different in its nature, content and origin into the centralized coordinated information resource. The data selection from primary sources, its fixation, filtering, transformation into the specific format to form content and set it in the database determine the information resources forming and usage order in a IRPIS (Fig. 5).

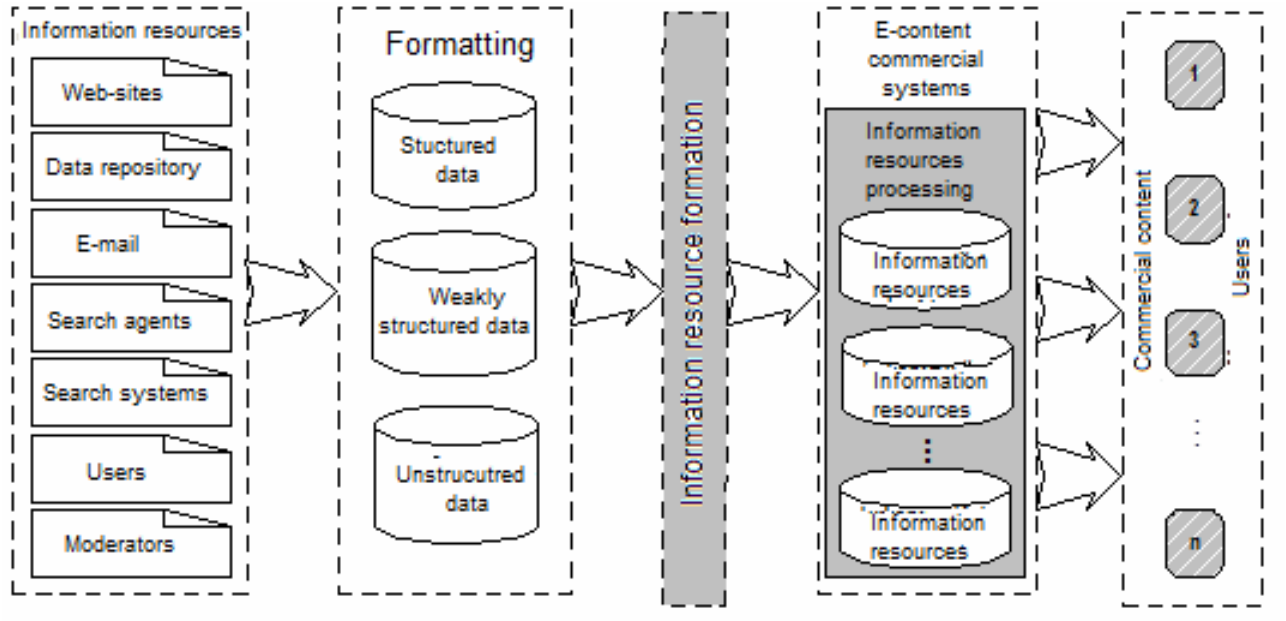


Fig. 5. The information resources forming and usage order in a IRPIS

The main IRPIS framing project task is the information resource architecture development by forming the actual commercial content which is created accordingly to the reaction of users on the commercial activity extension type (Fig. 6).

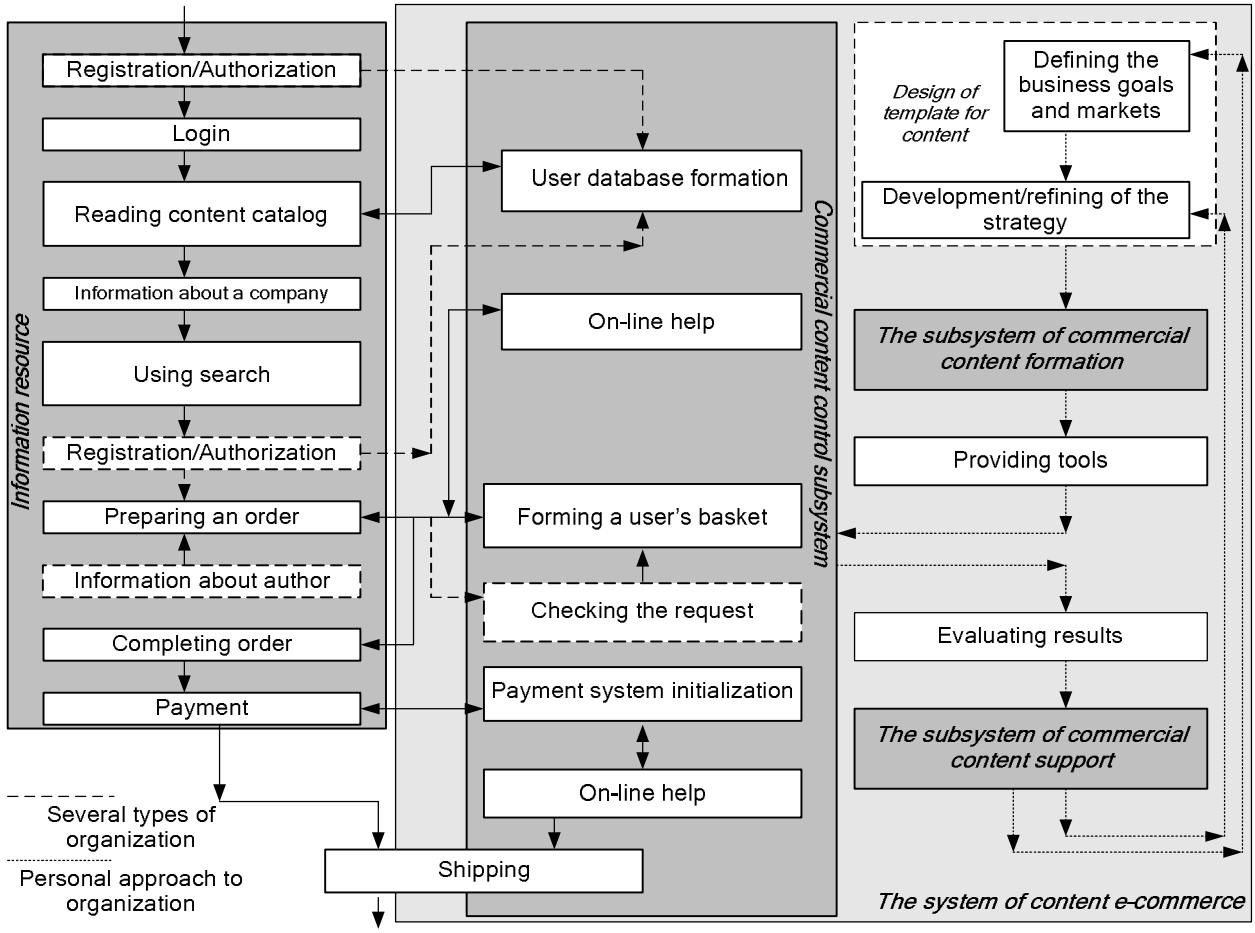


Fig. 6. The scheme of information torrents in e-content commerce systems

Let exist some prescribed initial content sources set n_X with fixed or variable composition. Every information source $Sourse(x_i)$, where x_i is the i-th content from the source under $i=\overline{1,n_X}$, forms some set of values which

contain statements/ knowledge /facts from the IRPIS subject field. The result of any requests to the source $Sourse(x_i)$ performed by IRPIS's technological means is the generating of the values set $X = \{x_1, x_2, ..., x_{n_X}\}$ that is acquired and presented in some fixed form. According to the system technological features, during the generated values selection and fixing process every values set generated by any of the information sources is transformed to the content input $C = \alpha(u_f, x_i, t_p)$ of the fixed format c_r, where $r = \overline{1, n_C}$.

Every set of content is presented the form of structured, weakly structured data or data without any specific structure description and is saved in the commercial content database $DataBase(C)$. For every single set the content structuring requires the forming of the set composition description, ways to combine elements and to normalize them (the set of terms $U = \{u_1, u_2, ..., u_{n_U}\}$ where u_f is the content forming condition under $f = \overline{1, n_U}$). The set of data from a source is the values set combination under certain format and set of terms $\langle X, U \rangle$ while forming the content input without any structure description ($U = \varnothing$). Before being saved the content should pass the verification / validation to confirm its formal / meaning correctness / relevance accordingly to the system requirements. If some material don't meet the specific measures this part of content is excluded from the further usage. The filtered content is formatted and saved and then the corresponding statements and knowledge become available for users via the- IRPIS i.e. $Sourse(x_i) \rightarrow x_i \in X \rightarrow X \rightarrow \langle X, U \rangle \rightarrow \alpha(u_f, x_i, t_p) \rightarrow c_r \rightarrow C \rightarrow DataBase(C) \rightarrow \beta(q_d, c_r, h_k, t_p) \rightarrow \langle C, H \rangle$ where $i = \overline{1, n_X}$, with n_X as the amount of content sources; $Sourse(x_i)$ – the source of i-th content; $x_i \in X$ – the i-th content of the source $Sourse(x_i)$; $X = \{x_1, x_2, ..., x_{n_X}\}$ – the set of data as a result of selection from the source $Sourse(x_i)$; $\langle X_i, U_i \rangle$ – the set of data with the terms set; $\alpha(u_f, x_i, t_p)$ – the content forming operator; c_r – the formed commercial content; C – formed content set; $DataBase(C)$ – the operator that saves commercial content in the database; $\beta(q_d, c_r, h_k, t_p)$ – the content managing operator; $\langle C, H \rangle$ is formed from the commercial content set and the informational resource content controlling terms.

The main stages of the information resource working out process are the content forming, control and maintenance with the following connections *content → content formation → database → content control → information resource / user request → content control → information resource → content maintenance → database*. The information resources working-out process may be presented as

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

where: X is the content stock from various resources, Q is the users' searches set, $Formation$ is the content forming operator, H is the forming and controlling terms set, C is the commercial content set, V is the terms set concerning the content maintenance and the external influence on the system, $Management$ is the content managing operator, $Support$ is the content maintenance operator, Z is the information resource component set, T is the time of information resource processing transactions, Y is the system functioning statistical data set. The commercial content forming operator provides the commercial content transformation into the new state which differs from the previous one by the new content portion that supplements the previous state. The commercial content managing operator provides the commercial content transformation into the new state that differs from the previous one by values of the distinguishing parameters (the actuality, completeness, relevance, authenticity, trustworthiness) which should answer some specific requirements. The commercial content maintenance operator provides the commercial content transformation into the collection of values which are created as the result of the analysis, monitoring and evaluation of the cooperation between a user, search systems and other informational resources that are the decision-making base for the content formation and management. The content formation phase is outlined in the $Formation$ operator like $c_r = Formation(u_f, x_i, t_p)$, with u_f as the set of content formation terms, i.e. $u_f = \{u_1(x_i), \ldots, u_{n_U}(x_i)\}$.

Thus the content is presented in the following way

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

The commercial content formation phase is outlined in the *Management* operator like $z_w = Management(q_d, c_r, h_k, t_p)$, where Q is the set of requests, H is the set of content managing terms, i.e. $H = \{h_1(c_{i+1}, q_d), \ldots, h_{n_H}(c_{i+n_H}, q_d)\}$. The content management process is presented as

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

The maintenance phase is described in the *Support* operator of the next format

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

where: v_l is the set of content managing terms and influences the environment has on the system i.e. $v_l = (v_1(q_i, h_k, c_r, z_w, t_p), \ldots, v_{n_V}(q_i, h_k, c_r, z_w, t_p))$. The output data is realized in the following way

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

Content formation is the data (from various information sources) processing control supporting measure complex for commercial content framing with the set of additional values such as actuality, authenticity, uniqueness, completeness, accuracy, etc. Content management is the determinant content parameters' (like actuality, completeness, relevance, authenticity, trustworthiness according to prescribed requirements in the criterion set) values supporting complex of measures.

Content maintenance is the complex of measures that provides the IRPIS's functioning in accordance with the prescribed requirements and the following ones. The complex system of related operations, methods and means (Fig. 7) is typical for any full-scale IRPIS. The content formation process may be presented by the following link pattern $Sourse(x_i) \rightarrow x_i \in X \rightarrow X \rightarrow \langle X, U \rangle \rightarrow \alpha_1(Downloading(\langle X, U \rangle), T) \rightarrow \alpha_2(Verification(\langle X, U \rangle), T) \rightarrow \alpha_3(Conversion(\langle X, U \rangle), T) \rightarrow \alpha_4(\langle X, U \rangle, T) \rightarrow \alpha_5(Qualification(\langle X, U \rangle), T) \rightarrow \alpha_6(\langle X, U \rangle, T) \rightarrow \alpha_7(\langle X, U \rangle, T) \rightarrow c_r \in C$,

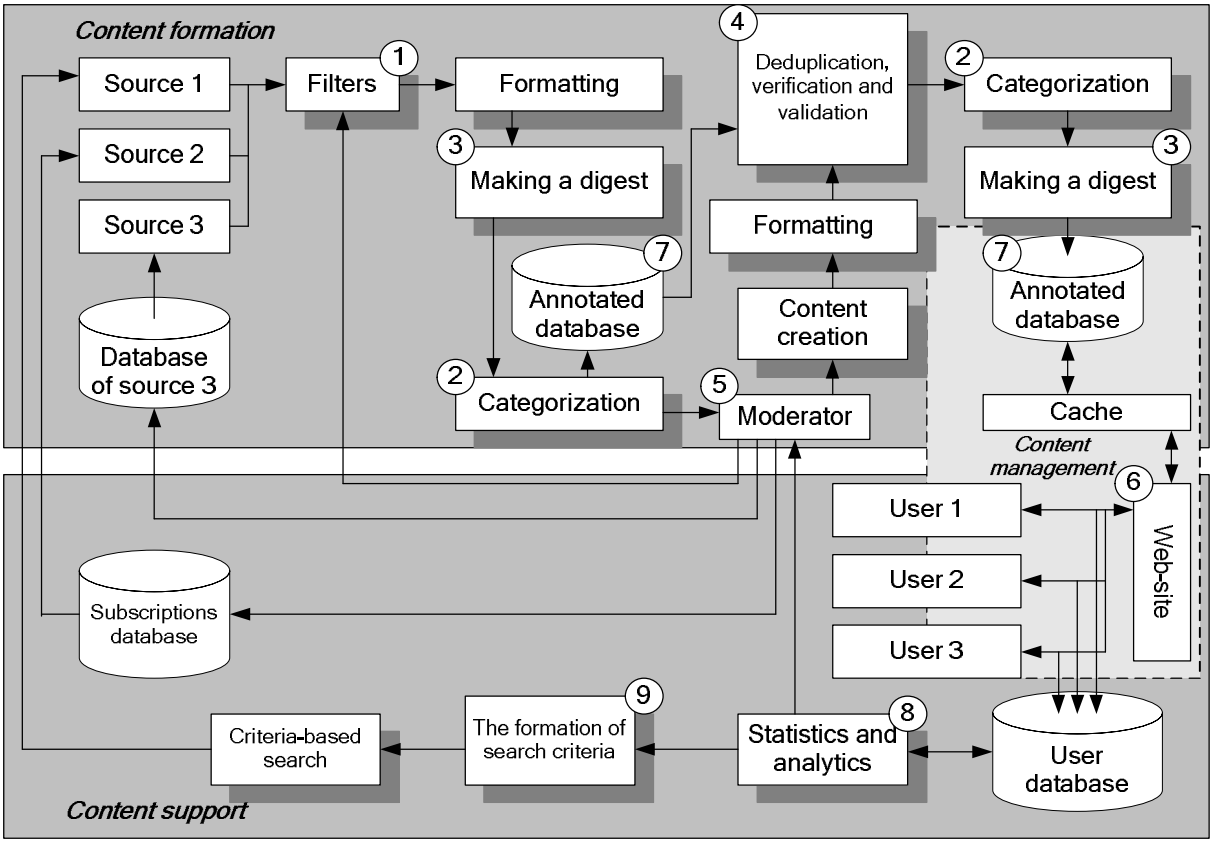


Fig. 7. The information resources processing methods

where: $X = \{x_1, x_2 \ldots, x_{n_X}\}$ is the input data $x_i \in X$ from various information resources or moderators under $i = \overline{1, n_X}$; α_1 is the operator that provides the content collecting from various resources; α_2 is the content reduplication detecting operator; α_3 is the content formatting operator; α_4 is the content's key words and concepts detecting operator; α_5 is the content automatic heading operator; α_6 is the content digest forming operator; α_7 is the selective content dissemination operator; $T = \{t_1, t_2 \ldots, t_{n_T}\}$ is the commercial content forming transaction's time $t_p \in T$ under $p = \overline{1, n_T}$; $C = \{c_1, c_2 \ldots, c_{n_C}\}$ is the commercial content set $c_r \in C$ under $r = \overline{1, n_C}$; $Verification(\langle X, U \rangle)$ is the content verification operator, $Qualification(\langle X, U \rangle)$ is the content qualification operator; $Conversion(\langle X, U \rangle)$ is the content conversion operator; $Downloading(\langle X, U \rangle)$ is the content downloading operator.

The process of commercial content formation for the information resource provides the link between the set of input data from various data sources and the formatted and saved commercial content set $S(x_i) \rightarrow x_i \rightarrow X \rightarrow \alpha(u_f, x_i, t_p) \rightarrow c_r \rightarrow C \rightarrow D(C)$, with $S(x_i)$ as the data resource, $S(x_i)$ as the commercial content database.

The content resources' types for the content formation subsystem are the following: the address list of information resources with confidential and other required

data; the address list of information resources with content subscription; the content set received from content moderators and authors; the request list with keywords for search systems. The content formation subsystem provides the gathering of information from various sources, its formatting, keyword detection, digest doubling and formation, heading and selective content dissemination. The commercial content formation subsystem is realized in the pattern of the content monitoring complexes which collect content from various data sources that provides the content database creation according to the customers' informational requests. In the result of the collecting process and pre-processing the content comes to the unique format, gets classified according to the fixed heading option, the descriptors with keywords becomes ascribed to the content. That makes the commercial content management process much simpler. The content managing system has the following tasks: the formation and rotation of the databases, providing access to them, the formation of the operative and retrospective databases, user work personalization, the retention of the personal users' requests and sources, functioning statistics conducting, the providing of the search within the databases, the output forms generation, the information interplay with other databases, the information resource formation. The commercial content managing subsystem is realized with the cashing usage (the presentation subsystem generated page only once, hereafter it loads several times faster from the cash which updates automatically after some period of time, when some parts of the informational resource was changed or manually by the administrator's request) or using the informational blocks (blocks are saved when the information resource is being edited; the page gets constructed of these blocks because of the user's request to open the corresponding page). The content maintenance subsystem provides the informational images formation; the content's thematic subject detection; the content interrelation table construction; the content ratings calculation; the detection of new events in the content streams, their monitoring and clustering.

CONCLUSIONS

The distribution is normally performed by moderators. The content distributional subsystem shortens the time and diminishes the resource usage for the further IRPIS functioning. The distributional process expects several stages to be processed: the distributional objects list formation (e.g. articles, software, books or digests); the estimation of the content distributional criterions / features from the received list (the content uniqueness percent, the amount of the content requests, the users' mark and the review time); the content authors ratings creation; the content's parameters evaluation for the purpose to use it within the distributional process. The listed criterions are not of the same importance and significance within their total analysis and the computing of the content authors' work quality composite mark. The content contains the thematics and the digest. The content distributional system selectively sends digests to authors according to their work quality ratings. The increase in the content volume causes the proper quality and productivity evaluation of the very content author. The increase in the evaluation criterions allows covering the broader aspect range of any author's / moderator's work.

REFERENCES

1. **Baeza-Yates R. and Rebeiro-Neto B. 1999.** Modern Information, Menlo Parl, California, New York : ACM Press, Addison-Wesley. Available online at: <http://people.ischool.berkeley.edu/~hearst/irbook/print/chap10.pdf>.
2. **Boiko B. 2004.** Content Management Bible. Hoboken.
3. **Braslavski P. and Tselishchev A. 2005.** Style-Dependent Document Ranking. In Proc. RCDL'2005. Available online at: <http://www.rcdl2005.uniyar.ac.ru/ru/RCDL2005/papers/sek7_1_paper.pdf>.
4. **Brin S. and Page L. 2005.** The Anatomy of a Large-Scale Hypertextual Web Search Engine. Available online at: <http://www-db.stanford.edu/pub/papers/google.pdf>.
5. CM Lifecycle Poster. **Content Management Professionals**. **2010.** Available online at: <http://www.cmprosold.org/resources/poster/>
6. **EMC, IBM and Microsoft. 2008.** Content Management Interoperability Services. Part I. Version 0.5. Hopkinton, 76.
7. **Grefenstette G. 1995.** Automatic Thesaurus Generation from Raw Text using Knowledge-Poor Techniques. Proceedings of SIGIR.
8. **Hearst M.A. 1992.** Automatic Acquisition of Hyponyms from Large Text Corpora. Proc. of the 14th International Conference on Computational Linguistics, Nantes, France, Available online at: <http://acl.ldc.upenn.edu/C/C92/C92-2082.pdf>.
9. **Karlgren J. and Cutting D. 1994.** Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In Proceedings of the 15th International Conference on Computational Linguistics, Kyoto, vol. 2, Pp. 1071-1075. Available online at: <http://www.sics.se/~jussi/Papers/1994_Coling_Kyoto_l/cmplglixcol.ps>.
10. **Manning C.D., Schutze H. 2005.** Foundations of Statistical Natural Language Processing. Chapter 5: Collocations. E-version of the chapter.
11. **Manning D.C., Raghavan P., Schutze H. 2007.** Introduction to Information Retrieval / // Cambridge University Press. The upcoming book's chapters.
12. **Manning D.C., Raghavan P., Schütze H. 2008.** An Introuction to information retrieval. Cambridge University Press, Cambridge, England. Available online at: <http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>.
13. **Sato S., Sasaki Y. 2003.** Automatic collection ofrelated terms from the web. In Proc. 41st ACL, Pp.121–124.
14. **Sebastiani F. 2002.** Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol. 34, No. 1, Pp.1-47.
15. **Segalovich I. 2002.** A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine, Available online

at: <http://company.yandex.ru/articles/iseg-las-vegas.html + the forum discussion>.

16. **Takkinen J. 2006.** IRI: Introduktion och IR-systemgrunder (modellering och utvardering). IISLAB/ADIT/IDA, Linkopings universitet. Available online at: <https://www.ida.liu.se/~TDDC08/tddc08-ir1.pdf>.
17. **Zamir O. and Etzioni O. 1998.** Web Document Clustering: A Feasibility Demonstration. In Proc. SIGIR'98.
18. **Ayvazyan S.A. 1989.** Prikladnaya statistika: Klassifikatsiya i snizhenie razmernosti: Sprav. izd. Under S.A. Ayvazyan redaction. M.: Finansyi i statistika.
19. **Berko A., Vysotska V., Pasichnyk V. 2009.** Electronic content commerce systems. Lviv: NULP, p.612. (in Ukrainian).
20. **Braslavskiy P.I. 2006.** Intellektualnyie informatsionnyie sistemyi. Available online at: <http://www.kansas.ru/ai2006/>.
21. **Braslavskiy P.I., Vovk E.A. and Maslov M.Yu. 2002.** Fasetnaya organizatsiya internet-kataloga i avtomaticheskaya zhanrovaya klassifikatsiya dokumentov. Kompyuternaya lingvistika i intellektualnyie tehnologii. tretiy internatsionalniy seminar "Dialog-2002". T. 2. Moscow.: Nauka. Pp. 83-93. Available online at: <http://company.yandex.ru/articles/article8.html>.
22. **Braslavskiy P.I., Kolyichev I. 2005.** eXtragon: eksperimentalnaya sistema dlya avtomaticheskogo referirovaniya veb-dokumentov. Trudyi ROMIP-2005. Saint Petersburg. Pp. 40-53. Available online at:<http://www.romip.narod.ru/romip2005/03_extragon.pdf>.
23. **Gavrilova T.A., Chervinskaya K.R. 1992.** Izvlechenie i strukturirovanie znaniy dlya ekspertnyih sistem. M.: Radio i svyaz. (in Russian).
24. **Gavrilova T.A., Horoshevskiy V.F. 2000.** Bazyi znaniy intellektualnyih sistem, Saint Petersburg. (in Russian).
25. **Gladkiy A.V. 1985.** Sintaksicheskie strukturyi estestvennogo yazyika v avtomatizirovannyih sistemah obscheniya. Moscow.: Nauka. (in Russian).
26. **Dobrov B.N., Lukashevich N.V., Syiromyatnikov S.V. 2003.** Formirovanie bazyi terminologicheskih svlovosochetaniy po tekstam predmetnoy oblasti. Elektronnyie biblioteki: Trudyi konferentsii RCDL Saint Petersburg. 201-210. Available online at: <http://rcdl2003.spbu.ru>.
27. **Dobryinin V. 2002.** Teoriya informatsionno-logicheskih sistem. Informatsionnyiy poisk. Saint Petersburg. Available online at: <http://ir.apmath.spbu.ru/publications/dobrynin_ir_intro/>.
28. **Ivanov V. 1994.** Kontent-analiz: Metodolohiia i metodyka doslidzhennia ZMK. Kyiv. p.112. (in Ukrainian).
29. **Ivanov S., Krukovskaya N. 2004.** Statisticheskiy analiz dokumentalnyih informatsionnyih potokov Nauchno-tehnicheskaya informatsiya. No 2. Pp.11-14. (in Russian).
30. Iskusstvennyiy intellekt: Spravochnik: **Kn.1: Sistemyi obscheniya i ekspertnyie sistemyi**. **1990.** Moscow: Radio i svyaz. (in Russian).
31. Iskusstvennyiy intellekt: Spravochnik: **Kn.2: Modeli i metodyi. 1990.** Moscow: Radio i svyaz, (in Russian).
32. **Clifton B. 2009.** Google Analytics: professional attendance analysis web sites. Moskva: Williams, p.400 (in Russian).
33. **Kovalenko A. 2006.** Veroyatnostnyiy morfologicheskiy analizator russkogo i ukrainskogo yazyikov. Available online at: <http://www.keva.ru/stemka/stemka.html>.
34. **Kukushkina O.V., Polikarpov A.A., HmelYov D.V. 2001.** Opredelenie avtorstva teksta s ispolzovaniem bukvennoy i grammaticheskoy informatsii. Problemyi peredachi informatsii, T.37, No.2. Pp.96-108. Available online at: <http://www.math.toronto.edu/dkhmelev/PAPERS/published/gramcodes/gramcodes.pdf>.
35. **Lande D., Furashev V., Braychevskyy S. and Grigoriev A. 2006.** Modeling and evaluation electronic information streams fundamentals. Kyiv: Engineering, p.348. (in Ukrainian).
36. **Leonteva N.N. 2006.** Avtomaticheskoe ponimanie tekstov: sistemyi, modeli, resursyi. Moscow: Izdatelskiy tsentr "Akademiya". (in Russian).
37. **Neyl K. and Shanmagantan G. 2005.** Web-instrument dlya vyiyavleniya plagiata. Otkryityie sistemyi. #01. Pp.40-44. Available online at: <http://www.osp.ru/os/2005/01/040_print.htm>.
38. **Nekrestyanov I., Panteleeva N. 2002.** Sistemyi tekstovogo poiska dlya Veb. Programmirovanie. No28(4). Pp.207-225. Available online at: <http://meta.math.spbu.ru/~nadejda/papers/web-ir/web-ir.html>.
39. **Osuga S. 1989.** Obrabotka znaniy. Moscow: Mir. (in Russian).
40. **Vysotska V., Sherbyna Y., Pasichnyk B., Shestakevich T. 2012.** Mathematical linguistics. Lviv: "Novy Svit - 2000", p.359.
41. **Penrouz R. 2003.** Novyiy um korolya: O kompyuterah, myishlenii i zakonah fiziki. Moscow: URSS. (in Russian).
42. Perspektivyi razvitiya vyichislitelnoy tehniki v 11 kn. **Kn. 2. Intellektualizatsiya EVM. 1989.** Moscow: Vyisshaya shkola. (in Russian).
43. **Popov E.V. 1982.** Obschenie s EVM na estestvennom yazyike. Moscow: Nauka. (in Russian).
44. **Popov E.V., Fominyih I.B., Kisel E.B., Shapot M.D. 1996.** Staticheskie i dinamicheskie ekspertnyie sistemyi. Moscow: Finansyi i statistika. (in Russian).
45. **Rao S.R. 1968.** Lineynyie statisticheskie metodyi i ih primeneniya. Moscow: Nauka. (in Russian).
46. **Segalovich I.V.** Kak rabotayut poiskovyie sistemyi / I.V. Segalovich // Mir Internet, - 2002. - #10. Available online at: <http://www.dialog-21.ru/directions/Segalovich_vorprint.doc.
47. **Sokirko A.V.** Morfologicheskie moduli na sayte www.aot.ru / A.V. Sokirko //Materialyi konferentsii "Dialog-2004". Available online at:

<http://www.dialog-21.ru/Archive/2004/Sokirko.htm.

48. **Solton D. 1979.** Dinamicheskie bibliotechno-informatsionnyie sistemyi.Moscow: Mir, p. 560. (in Russian).
49. **Faktornyiy, diskriminantnyiy i klasternyiy analiz. 1989.** Moscow: Finansyi i statistika. (in Russian).
50. **Fedorchuk A. 2005.** Content Monitoring information flows. Nat. Acad. Science Problems. Functioning, Trends of development. Vol. 3. Available online at: <http://www.nbuv.gov.ua/articles/2005/05fagmip.html>
51. **Han U., Mani I. 2000.** Sistemyi avtomaticheskogo referirovaniya. Otkryityie sistemyi. No12. Available online at: <http://www.osp.ru/os/2000/12/067_print.htm>.
52. **Hmelev D. 2000.** Raspoznavanie avtora teksta s ispolzovaniem tsepey A.A. Markova. Vestnik MGU, S.9: Filologiya, No2, Pp.115-126. Available online at: <http://www.rusf.ru/books/analysis/vestnik2000win>.htm.
53. **Hramtsov P. 1996.** Informatsionno-poiskovyie sistemyi Internet. Otkryityie sistemyi. No3. Available online at: <http://www.osp.ru/os/1996/03/46_print.htm>.
54. **Lytvyn V. 2013.** Design of intelligent decision support systems using ontological approach. Econtechmod: Lublin, Rzeszow, Vol. II, No 1, Pp.31-38. (in Poland).
55. **Lytvyn V., Semotuyk O., Moroz O. 2013.** Definition of the semantic metrics on the basis of thesaurus of subject area. Econtechmod: Lublin, Rzeszow, Vol. II, No 4, Pp.47-51. (in Poland).
56. **Vysochina M. 2014.** The innovative approach to the study of decision-making in the context of the specific character of a product of managerial work. Econtechmod: Lublin, Rzeszow, Vol. III, No 2, Pp.87-92. (in Poland).
57. **Rybytska O., Vovk M. 2014.** An application of the fuzzy set theory and fuzzy logic to the problem of predicting the value of goods rests. Econtechmod: Lublin, Rzeszow, Vol. III, No 2, Pp.65-69. (in Poland).
58. **Fedasyuk D., Yakovyna V., Serdyuk P., Nytrebych O. 2014.** Variables state-based software usage model. Econtechmod: Lublin, Rzeszow, Vol. III, No 2, Pp.15-20. (in Poland).
59. **Ryshkovets Yu., Zhezhnych P. 2013.** Information model of Web-gallery taking into account user's interests. Econtechmod: Lublin, Rzeszow, Vol. II, No 3, Pp.59-63. (in Poland).
60. **Vysotska V., Rishnyak I., Churun L. 2007.** Analysis and evaluation of risks in electronic commerce. CAD Systems in Microelectronics, CADSM '07, 9th International Conference. Pp.332-333. (in Ukrainian).