computer linguistic system, texts analysis and synthesis, content system management

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

FEATURES OF THE CONTENT-ANALYSIS METHOD FOR TEXT CATEGORIZATION OF COMMERCIAL CONTENT IN PROCESSING ONLINE NEWSPAPER ARTICLES

Abstract

This paper presents the features of text categorization of commercial content in linguistic modelling. Description of syntax sentence modelling is applied to automate the processes of analysis and synthesis of texts in natural language for commercial content categorization. This article suggests methods of content analysis for online newspaper. The model describes the processing of information resources systems of content analysis and simplifies the technology of content management system automation. General problems of syntactical and semantic content analysis and functional services of content management system are analysed.

1. INFORMATION

The methods and tools development for automatic processing of text of commercial content in modern information technology are important and topical [1–5] (for example, systems of information retrieval, machine translation, semantic, statistical, optical and acoustic analysis and synthesis of speech, automated editing, knowledge extracting from the text content, text content abstracting and annotation, textual content indexing, training and didactic, linguistic buildings management, instrumental means of dictionaries conclusion of various types, etc.) [6–15]. Specialists actively seeking new models of description and methods for automatic processing of text content [2–4]. One of these methods is the development of general principles of lexicographic systems of syntactic type. It is important by these principles these systems construction of text content processing for specific languages [1, 5].

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

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

In the last ten years humanity has implemented a significant step in developing and implementing new technologies. Development of technologies has given the opportunity to solve a lot of complex tasks, which touch humanity, but also generate new tasks, solution of which is difficult. One of these tasks is a task of content analysis. Methods and systems of content analysis are used in various areas of human activity (politics, sociology, history, philology, computer science, journalism, medicine, etc.) [1–5]. These systems are quite successful and do not require large funds and time to get the desired result. At the same time using this type product allows you to increase the level of success at 60 %. Basic system of content analysis includes the following features: quick information updates, searching for information on this resourse, collect data about the customers and potential customers, creating and editing surveys, analysis of resource visitations. If to automate system for the using information system of content analysis, the workload can be reduced, the time for processing and obtaining the necessary information can be also reduced, productivity of work system increases which leads to a decrease in expenses of money and time to get the desired result. Issue of the theme has been caused by increasing demands of the users of these systems and by the following factors: rapid growth in demand for reliable information, the necessity of forming plurals operational information as well as use for automatic filtering unwanted information [1–5].

2. RECENT RESEARCH AND PUBLICATIONS ANALYSIS

Any tools of syntactic analysis consists of two parts: a knowledge base about a particular natural language and algorithm of syntactic analysis (a set of standard operators of text content processing on this knowledge) [1–5]. The source of grammatical knowledge is data from morphological analysis and various filled tables of concepts and linguistic units [2]. They are the result of the empirical processing of textual content in natural language of experts in order to highlight the basic laws for syntactic analysis. Table-based of linguistic units constitute configurations or valences sets (syntactic and semantic-syntactic dependencies) [2]. This is a lexical units list/dictionaries as instructions for every of them all possible links with other units of expression in natural language [2, 5]. In implementing of the syntactic analysis should be achieved full independence of rules of tables data transform from their contents. This change of this content does not require algorithm restructuring.

The vocabulary V consists of finite not empty set of lexical units [2]. The expression on V is a finite-length string of lexical units with V. An empty string does not contain lexical items and is denoted by Λ. The set of all lexical units over V is denoted as V'. The language over V is a subset V'. The language displayed through the set of all lexical units of language or through definition

criteria, which should satisfy lexical items that belong to the language [2]. Another is one important method to set the language through the use of generative grammar. The grammar consists of a lexical units set of various types and the rules or productions set of expression constructing. Grammar has a vocabulary *V*, which is the set of lexical units for language expressions building. Some of lexical units of vocabulary (terminal) can not be replaced by other lexical units.

3. RESEARCH RESULTS ANALYSIS

Development of Internet technologies and its services gave the humanity access to virtually unlimited quantity of information but as often happens in these cases – there is a problem in reliability and efficiency. It is for that, because the information was efficient and trustworthy, technology of content analysis are implemented. The use of these technologies allows you to receive the information as a result of her functioning, provides an opportunity to interference in the system operation to increase the level of that system, the activity of the information resource and for popularity increase among the users. World's leading producers of processing information resources work actively in this direction such as Google, AIIM, CM Professionals organization, EMC, IBM, Microsoft alfresco, Open Text, Oracle, SAP.

Content analysis is a high-quality and quantitative method information studies, which is characterized by objectivity of conclusions and austerity of procedure and is in the quantitive treatment of results further interpretation [1]. Content Management System, CMS is a software for web-sites organization or other information resources in the Internet or computer networks [1]. Today there are hundreds of available CMS and due to the functionality they can be used in different areas. Despite the wide range of tool and technical facilities available at CMS properties for all content management systems are similar. The Web content management system (WCMS) is a software complex which provides functions of creating, editing, control and organization Web pages. WCMS is often used for blogs creation, personal web pages and online-shops and are intended for users, who are not familiar with programming [1].

The following analysis stages are identified [5]:

1. *Program preparation for the document analysis.* At this stage so-called empirical theory research is being formulated. That means in analysis preparation hypotheses, which are in the context of problems are being systematised and those are discarded these that are not exposed to the data information.
2. *Selection of analysis sources.* It is necessary to identify sources, which include materials and information.

3. *Analysis of empirical models, the selection* (human communication, choice of materials for different periods of time, the types of messages, type of selection).
4. *Development of the specific analysis methods.*
5. *Platane research, testing of methods reliability*.
6. *Collection of primary empirical information*.
7. *Quantitative processing of collected data*.
8. *Interpretation of results, research conclusions*.

4. METHODS OF CONTENT ANALYSIS

Content analysis is based on journalism and mass communication and uses equipment in the following empirical areas: psychiatry, psychology, history, anthropology, education, philology and literature analysis, linguistics. Overall, the methods of content analysis in these areas so or otherwise is connected with the use in the sociological research framework. Content analysis is rapidly developing today, it is associated with development of information and Internet technologies, where this method has found wide application.

While creating an effective information system significant attention should be given to content management, because content analysis is used in the content management systems for automation of work and lowers expenses of time and money. There are several stages in the content management such as: content analysis, the content processing and submission of content. For effective system work firstly, analyze the content, then process the relevant results and make conclusions and then work on the that content. And on the final step is the presentation of content. The following methods of content analysis are: comments analysis, rating evaluation, analysis of statistics and history [5].

Comments analysis is used for adjustment analysis of the system user's moods who write in its comments reviews about system advantages and disadvantages or for adjusting operational and liquid information.

Analysis of statistics and history is used for observation and result processing, which are used to determine information efficiency and liquidity. For example, if one of the articles was visited by 100 users and another by 1, then you can certainly say that the information is more efficient in first article than in the second [5].

Rating assessment is used to determine the rate the same articles and is conducted by the polls, the evaluation users, etc. (Fig. 1).

Graphic content analysis is caused by the fact that in most cases graphic information is absorbed users faster than text. This can be seen, for example, in the motion diagrams, charts and diagrams (in the form information will be assimilated slower). When applying content analysis text use appropriate methods, and in this case, use two types of analysis: a quantitative and qualitative.

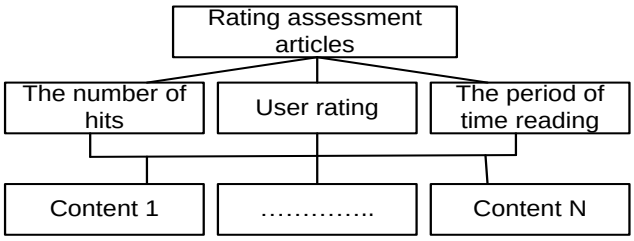

Fig. 1. Products rating assessment articles [source: own study]

A quantitative content analysis must include standardized procedures of counting allocated categories (table 1). For conclusions crucial role have quanti-tative values that are characterized by one or another category. Indicators may differ or, on the contrary, to be friends with an absolute value, which will be taken into consideration during interpretation processing results.

Tab. 1. Stages of quantitative content analysis

The stage	Stage characteristic
Allocation unit analysis	Conversion of linguistic unit in the form for processing.
Frequency units counting	Identify relationships between linguistic units.
Categorization	Determination of finite sets of redundant categories for obtaining quantitative data of their appearance.
Data mining	In-line detection content through multiple quantitative assessments of knowledge with further qualification of them as categories.
Interpretation of	Getting content and semantically- filled results using mathematical methods and semantic formers.

High-quality content-analysis aims to study text material deep and meaningful, as well as from the point of view of context in which the dedicated category (table 2). takes into account relationships of meaningful elements and their relative importance in the text structure. Depending on research tasks, high-quality content-analysis may be expanded with some quantitative content analysis elements [5].

Tab. 2. Stages of high-quality content-analysis

The stage	Stage characteristic
Wrapping text to blocks	Formation of integrated meaningful units for encoding and processing.
Reconstruction of flow content	The reconstruction of values, opinions, views and evidence each source text.
Conclusions formation	The generalizations through comparison of individual system values.

You can complicate the task, if put in as the precondition the allocation of substantial notional relation units of texts and then calculate the relative importance of the statements in comparison with other. In both cases the main part of calculations can be done with the use of computer programs [2]. The main stages of text information content-analysis:

1. *Determination of invistagated aggregate sources* or messages in accordance with the specified criteria, which corresponds to each message: the type of source (forum, e-mail, chat, online-newspaper, Internet); message types (article, electronic letter, banner, commentator); sides who take part in the communication process (the sender, the reciever, recipient); the messages size that they compare (the minimum amount of/length); frequency of messages apperance; distribution method; place of message distribution; time of the messages, etc.
2. *Forming a limited sample messages.*
3. *Identifying linguistic units of analysis*. There are strict requirements to a possible linguistic analysis unit: a big enough for interpretation importance; smallenough not to interpret many meanings; easily identified; the amount of units is large enough for the sample. In the case of the per unit analysis topics, take into account those rules: the theme does not go beyond the paragraph; a new topic is there when you change the theme, purpose, categories and persons, for which a topic is.
4. *Allocation of units calculation* which can coincide with meaningful units or have a specific character. In the first case, the analysis procedure is reduced to counting the frequency of the selected content items, in the other – a researcher from the analysed material and research purposes shows calculating units, which can be: physical text length; text area filled with informative units; number of lines (paragraphs, trademarks, text columns); the size and type of file, the amount of drawings with a content, plot, and so on. In some cases researchers use also other calculating elements. A fundamental value at this stage of content-analysis is a strict definition of its operators.
5. *Direct procedure for calculating* . In general terms, it is similar to standard methods of classification for selected groups of mathematical statistics and probability theory formulas. There are also special procedure about content-analysis counting.
6. *The received results interpretation to* a specific goals and tasks of research. At this stage characteristics of text are detected and evaluated and which allow to draw conclusions about what author wanted to emphasize/hide.

Content analysis – this is a high-quality and quantitative method to study the information, which is characterized by objectivity and strict procedures and is followed by the elaboration of quantative results interpretation. Because of significant step in developing and implementing new technologies mankind has

implemented in the last few decades, it created new tasks that are difficult to be solved. One of these tasks is to provide users with reliable and updated information. *Efficiency* – the information property, which means that gathering and processing of information follows the dynamics ofchanging the situation. *Reliability* – the property of information to be properly understood, the probability of no errors, unquestionable loyalty of the following information which takes a person. Thus, the reliability is not the same thing as the truth is. Details can be reliable or unreliable for those who perceive them, but not at all [5]. Content is the basis for online newspapers, where the user is looking for all the information he needs. That is why we need the information to be operational and available for the user. For example, if the text will have a large number of definitions and terms or formulas, it will be hard -readable and user will be uninterested in reading it. On the other hand, if the text will contain a large number of unnecessary information user will spend more time on reading. There is a need to solve actual scientific problems of automatic processing of online newspapers content to receive operational and available for the user information with information noise liquidation and less time needed on the formation process of the final data search result.

5. TEST REVIEW

Rating of the clauses is done with the help of 3 criteria. The amount of referencing, reading time, and custom views are taken into account. These criteria compose rating of the clauses (Fig. 2).

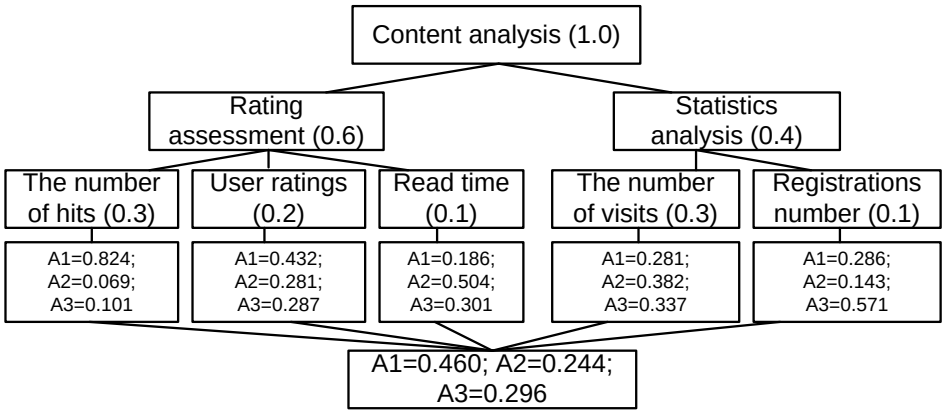

Fig. 2. Hierarchy of the choices for content analysis [source: own study]

Fig. 2 shows hierarchy of choices for content analysis. It consists of 3 levels: aim, criteria of 2 layers and alternatives. Congruous numbers of the fig 3 approve higher-priority elements of the hierarchy that are featured up with the help of MAI based on pairwise data comparisons on each layer related to elements of higher layer. The priorities of alternatives are calculated at the final stage by linear local priorities of all elements. Well-known higher layer criteria priorities pay the greatest attention to *Rating of the clauses* as it objectively reflects the quality of work. A statistic analysis criterion is less important because it doesn't reflect completely the quality of work. In the lower layer criteria attention is paid to *Amount of referencing* and visits that reflect users interest in a particular material. The next important criterion is a *Custom view* that shows the users' estimation for a specific material. The least vital criteria are *Reading time* and *Amount of signs up*. All the choices are figured separately due to each criteria. Due ot accounts it appears is the best for chosen criteria. An important factor for the functioning of the system is the availability of input data. Article is a journalistic or scientific work, which thoroughly and deeply, with scientific precision treats, interprets and summarizes the theoretical problem of social reality. Adding, editing and deleting of the articles is provided by an administrator. The source of information for articles are encyclopedias, periodicals, books, other articles, etc.. Another factor is the attendance of users. Users – are individuals who use information resource searching articles, reading them and voting? No user aspect impossible test quality content. Without user aspect, test content quality is impossible. There are two players in our system: *User* and *System manager*. 1st player is aimed for user modeling system as an individual. Player signs in / authorizes in the system, votes for the revised articles, reviews and searches for the articles. *System Manager* – is a person who analyzes and classifies information, adds edits and deletes articles, as well as performs other actions. He manages our system. Administration includes removal of incorrect articles and adding of new ones. *Sign in* option is used for registration of users in the system. *Authorization* option is used for user authentication. Usage of Authorization and *sign in* options expand the usage *statistic analysis*.

The object of *statistic analysis* is used for statistical estimation. Information is sent from the database in order to estimation. After the evaluation results are sent to the administrator object to this object could draw conclusions that would be needed to add new items or remove unwanted. After this procedure results are sent to the *system manager* so that he could conclude that what would be needed to add new items or remove unnecessary. System manager is the only person who can add or remove articles. After adding or deleting items the information is sent to display the latest and popular articles. Also, he checks for new articles, if the data was not sent. The result are new articles transmitted for publication. This object takes the output of new articles. This object performs publication of new items. On constructing cooperation graphs was held the determination of system objects, interaction analysis. Due to the graph there are 15 cocurrent flows.

Each flow put across items and informatiom from one object to another. Figure 3 shows working graph. It lets implement features of procedural synchro programming. Here you can see the system status after the user authentication. Fig. 3. shows that after the authorization user goes to browsing of the articles and the system shows him all the available items.

After the *browsing* the system shows user popular and latest articles. The user choose a specific item and the *choice of article* option helps to do that. After that the user goes to *Reading* option. This main option because here the user read the article and receive the information about the status of it. After the *reading* option the system goes into the following two statuses. These are *publication* status and *rating of the clauses*, that has 3 sub states. This is the *escape of amount of referencing* status. Here the information on the *amount of referencing* to this article is changed. The next step is is the *reading time change* status, where the time of reading the article is changed. The next one is *custom views* status where the the user rates for a particular article. These 3 statuses are used for the rating of the clauses. Afterthat the system goes to the *article shut-off*. This status is characterized by the persistence of data and results. The system goes into the final state. Fig. 4 presents content analysis process.

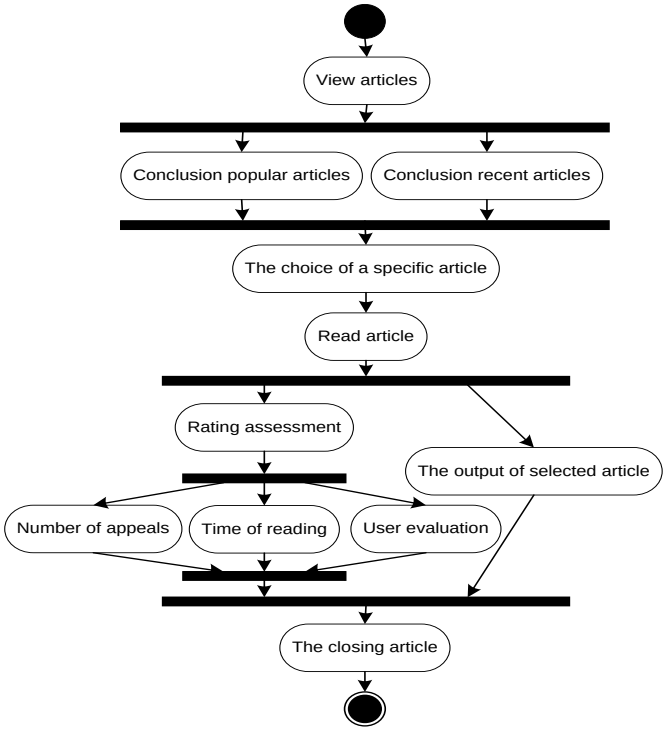

Fig. 3. Working graph [source: own study]

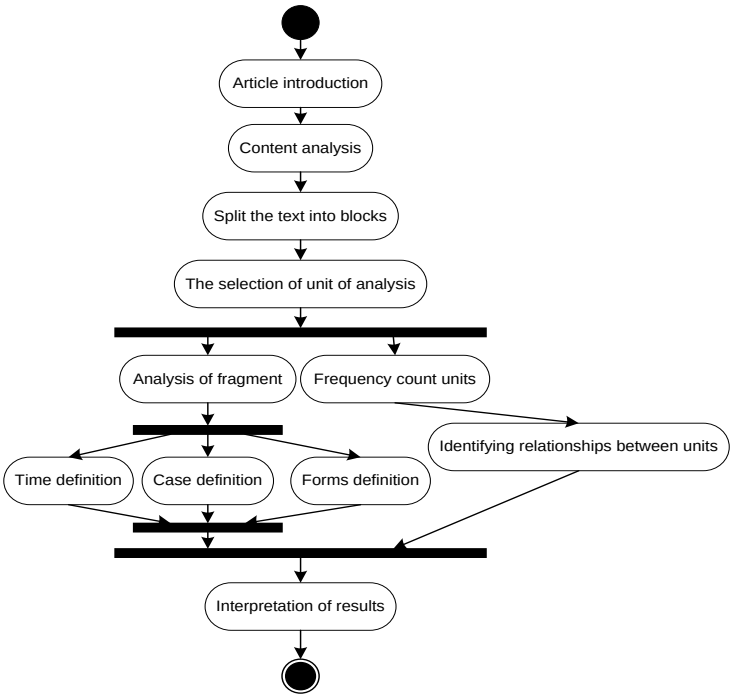

Fig. 4. Working graph of the content analysis [source: own study]

the publication of the article system enter *content analysis* status, after it goes to the next stage of *dividing the text into bodies*. Here the text is divided into items and the unit of analysis is identified. Farther fragment analysis is performed, counting of rate units, and identifying of relationships between linguistic units. After each option is held, the interpretation of the results of the content analysis is performed. Fig. 5 shows the working graph of the process of statistical analysis. Statistical analysis function after the system capture the data. One of the main options is *calculation of the amount of system users*. In addition. Amount of authentications and signs up are calculated. Then system calculates reading time of the articles and the average reading time. After all options are done, the treatment of statistical analysis is held.

The text realizes structural submitted activities through provides subject, object, process, purpose, means and results that appear in content, structural, functional and communicative criteria and parameters. The units of internal organization of the text structure are alphabet, vocabulary (paradigmatics), grammar (syntagmatic) paradigm, paradigmatic relations, syntagmatic relation, identification rules, expressions, unity between phrasal, fragments and blocks. On the compositional level are isolated sentences, paragraphs, sections, chapters, under the chapter, page etc. that (except the sentence) indirectly related to the internal structure because are not considered.

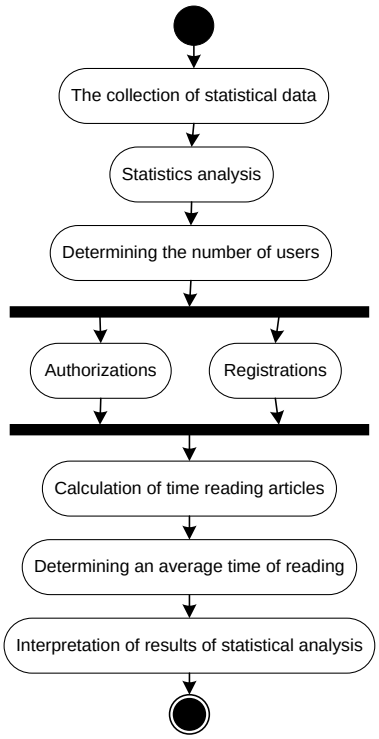

Fig. 5. Working graph of the statistical analysis [source: own study]

Content analysis for compliance thematic requests to $C_{Ct} = Categorization(KeyWords(C,U_K),U_{Ct})$, where $KeyWords(C,U_K)$ is the keywords identify operator, $Categorization$ is content categorize operator according to the keywords identified, U_K is keywords identify conditions set, U_{Ct} is categorization conditions set, C_{Ct} is rubrics relevant content set. Digest set C_D formed by such dependence as $C_D = BuDigest(C_{Ct},U_D)$, where $BuDigest$ is digests forming operator, U_D is conditions set for the digests formation, C_{Ct} is rubrics relevant content set. With the help of a database (database for terms/morphemes and structural parts of speech) and defined rules of text analysis searching terms. Parsers operate in two stages: lexemes content identifying and a parsing tree creates (alg. 1).

Algorithm 1. Parser of textual commercial content.
Stage 1. Content C lexemes identification from set *V*.
Step 1. Terms string definition over *V* as a sentence.
Step 2. Nouns groups identification with bases dictionary V'.
Step 3. Verbs groups identification with bases dictionary V'.

Stage 2. Creating of a parsing tree from left to right. Each step of output is the deployment as one symbols of the previous string or it to others replacing, while other symbols are rewritten without change. It is obtained the component tree, or syntactic structure, if process of deployment, replacement or re-writing characters (*fathers*) connect lines directly with symbols that come out as a result of the deployment, replacement or re-writing (*descendants*).

Step 1. Nouns group deployment. Verbs group deployment.

Step 2. The implementation of syntactic categories of word forms.

Stage 3. The keywords set determination.

Step 1. The *Noun* terms determination (nouns, nouns word combinations or adjective with the noun) among the words set of commercial textual content.

Step 2. The *Unicity* uniqueness calculation for *Noun* terms.

Step 3. The *NumbSymb* value calculation (the characters number with no spaces) for *Noun* terms at $Unicity \geq 80$.

Step 4. The *UseFrequency* value calculation (frequency of keywords using). The *UseFrequency* frequency for terms with $NumbSymb \leq 2000$ is within the limits $(6;8]$%. Frequency for terms with $NumbSymb \geq 3000$ is within the limits $[2;4)$%. Frequency for terms with $2000 > NumbSymb < 3000$ is within the limits $[4;6]$%.

Step 5. The values calculation of *BUseFrequency* (the frequency of keywords using at the beginning in the text), *IUseFrequency* (the frequency of keywords using in the middle of the text), *EUseFrequency* (the frequency of keywords using at the end in the text).

Step 6. Comparison of values *BUseFrequency*, *IUseFrequency* and *EUseFrequency* for priorities definition. Keywords with higher values *BUseFrequency* have higher priority than keywords with a higher value *IUseFrequency*.

Step 7. Keywords sorting according to their priorities.

Stage 4. The database filling of search image for content, i.e. attributes of *KeyWords* (keywords), *Unicity* (the keywords uniqueness ≥ 80), *Noun* (term), *NumbSymb* (the number of characters without spaces), *UseFrequency* (frequency of keywords using), *BUseFrequency* (frequency of keywords using at the beginning in the text), *IUseFrequency* (the frequency of keywords using in the middle of text), *EUseFrequency* (frequency of keywords using at the end in the text).

Based on the rules of generative grammar perform term correction under the rules of its use in context. The sentence define action limits of punctuation marks and links. The text semantics is due communicative task of information transfer. The textual structure is determined by internal organization of textual units and their relationship regularities. While parsing the text drawn in a data structure (eg. tree which corresponds the syntactic structure of the input sequence, and is best suited for further processing). After analysis textual block and term

is synthesized a new term as a keyword of content topics by using base of terms and their morphemes. Next is synthesized terms for a new keyword formation by using base of structural parts of speech. The principle of keywords detection in content (terms) is based on Zipf's law. It is reduced to words choice with an average frequency of occurrence. The most used words are ignored through the stop-dictionaries. And the rare words do not include text. According a meaningful analysis of the content corresponds to the process grammatical data extraction from the word by grapheme analysis and the results correction of morphological analysis through the grammatical context analysis of linguistic units (Alg. 2).

Algorithm 2. The textual commercial content categorization

Stage 1. The division of commercial content on the blocks.

Step 1. Block presentation to the input of tree construction with commercial content blocks.

Step 2. New block creation in the blocks table.

Step 3. The newline characters accumulation.

Step 4. Checking on point availability before a newline character. If there is, then go to step 5. If do not, then begin the sequence saving in the table, the new block parsing and transition to step 3.

Step 5. Checking on availability of the end in the text. If the end of the text is, then go to step 6. If do not, then start the accumulated sequence saved in the table, the new block parsing and transition to step 2.

Step 6. Blocks tree getting on the output as a table.

Stage 2. The block division on sentence with structure preservation.

Step 1. The input is a table of blocks. The sentences table creation with link for field *ID_section* of *n*-to-1 type of blocks table.

Step 2. A new sentence creation in sentences table.

Step 3. The symbols accumulation to point, semicolon or newline character.

Step 4. Checking on availability of cuts. If the cut exists, then go to step 5. If do not, then start the sequence saving in the table, a new sentence parsing and transition to step 2.

Step 5. Checking on availability of the end in the text block. If the end of the text exists, then go to step 6. If do not, then begin a sequence saving in the table, new sentences parsing and transition to step 2.

Step 6. The sentences tree getting on the output as a table.

Step 7. Checking for the end of the text. If the end of the text exists, then go to step 8. If do not, then start the new block parsing and transition to step 1.

Step 8. The sentences tree getting on the output in the form of tables.

Stage 3. The sentences division for lexemes with indication of belonging to sentences.

Step 1. The lexemes table formation based on the sentences table with fields of *ID_lexemes* (unique identifier), *ID_sentence* (number equal to the code of the sentence with lexeme), *Lexemes_number* (number equal to the lexemes number in the sentence), *Text* (lexeme text).

Step 2. A sentence presentation to the input from the sentences table for parsing on lexemes.

Step 3. A new lexeme creation in the lexemes table.
Step 4. The symbols accumulation to point, spaces or end of a sentence and the seving in the lexemes table.
Step 5. Checking for the end of the sentence. If yes, then go to step 6. If not then accumulated sequence saving in the table, new lexeme parsing and transition to step 3.
Step 6. Conducting parsing based on data obtained on the output (Alg. 1).
Step 7. Conducting morphological analysis based on data obtained at the output.
Stage 4. The topics determination for the commercial content.
Step 1. The hierarchical structure construction of properties for each lexical unit with text that includes grammatical and semantic information.
Step 2. The lexicon formation with a hierarchical organization of properties types, where each type-descendant inherits and overrides the ancestor properties.
Step 3. Unification as a basic mechanism for constructing syntactic structures.
Step 4. The *KeyWords* set identification for the commercial content (Alh.1).
Step 5. The values set definition as *TKeyWords* (thematic keywords in the *KeyWords* set for commercial content), *Topic* (the theme for commercial content) and *Category* (commercial content category).
Step 6. The values set definition as *FKeyWords* (the frequency of keywords using in the textual commercial content) and *QuantitativeryTKey* (frequency of thematic keywords using in the textual commercial content).
Step 7. The values set definition as *Comparison* (the keyword using comparison with various topics). The values set calculation as *CofKeyWords* (coefficient of thematic content keywords), *Static* (coefficient of the statistical terms importance), *Addterm* (coefficient of the additional terms availability). Comparison of the content keywords set with the key concepts with topics. If there is a match, then go to step 9. If not, then go to step 8.
Step 8. A new category formation with a set of key concepts of the analyzed content.
Step 9. Assignment designated section of the analyzed commercial content.
Step 10. The values set calculation as *Location* (the coefficient of content location in the thematic section).
Stage 4. Filling the search images base for attributes as *Topic* (the theme of content), *Category* (content category), *Location* (the coefficient of content location in the thematic section), *CofKeyWords* (the coefficient of thematic keyword of textual content), *Static* (coefficient of statistical significance for terms), *Addterm* (the coefficient of the additional terms availability), *TKeyWords* (the thematic keywords), *FKeyWords* (the frequency of using keywords), *Comparison* (the keywords using comparison of the different themes), *QuantitativeryTKey* (frequency of thematic keywords using in the text of commercial content).

The process of categorization through automatic indexing component in commercial content is divided into sequential blocks: a morphological analysis, a syntactic analysis, a semantic and a syntactic analysis of the linguistic structures and meaningful writing variation in the textual content.

6. CONCLUSIONS

Content as articles is the base of online newspaper due to which the user is looking for the necessary information. Thanks to content analysis, the system owner can determine the reliability and efficiency of the information contained in the articles of online newspaper. With the help of this option you can determine the popularity of the newspaper and do some actions in order to augmenter number of users. General recommendations in architectural design of content analysis systems are developed which, however, differ from existing by more detailed stages and availability of information processing module resources, allowing an efficient and easy to handle information resources at system developer's stage.

In the thesis known methods and approaches to solving the problem of automatic processing of textual content and selected advantages and disadvantages of existing approaches and results in the field of the syntactic aspects of computational linguistics is considered.

In this paper the general conceptual framework of modeling inflectional processes of the text arrays creation is formed. The syntactic model and inflectional classification of lexical structure of sentences is proposed. Also in the theses lexicographical rules of syntactic type for automated processing of these sentences is developed.

The proposed technique allows to achieve the highest parameters of reliability in comparison with known analogues. They also demonstrate the high efficiency of applied applications in the linguistics construction of new information technologies and research inflectional effects in natural language. The work has practical value, since the proposed models and rules can effectively organize the process of creating a lexicographical systems of textual content processing of syntactic type.

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.

REFERENCES

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

[2] BOLSHAKOVA E., LANDE D., NOSKOV A., KLYSHINSKY E., PESKOVA O., YAGUNOVA E.: *Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i kompyuternaya lingvistika.* (in Russian). Moskva: MIEM 2011, p. 272.

[3] LANDE D., FURASHEV V., BRAYCHEVSKY S., GRIGOREV O.: *Osnovy modelirovaniya i otsenki elektronnyh informatsionnyh potokov*. (in Ukraine). Kyiv: Inzhiniring, 2006, p. 348.

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

[5] PASICHNYK V., SCHERBYNA J., VYSOTSKA V., SHESTAKEVYCH T.: *Matematychna linhvistyka.* (in Ukraine). Lviv: "Novyy Svit – 2000", 2012, p. 359.

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

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

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

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

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

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

[12] MCKEEVER S.: *Understanding Web content management systems: evolution, lifecycle and market*. (in USA). Industrial Management & Data Systems (MCB UP),Vol. 103, No. 9, 2003, pp. 686–692.

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

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

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