Запропоновано підхід до розроблення системи розуміння особистості через контент-аналіз інформаційних ресурсів. Використано модель Big-Five на основі користувацької поведінки в соціальних мережах. Для визначення психологічних диспозицій розроблено метод аналізу англомовних та україномовних постів. Система призначена для рекрутингу, маркетингу, соціальних мереж і Web-сервісів. Аналіз аудиторії покращує ефективність контекстної реклами, систем рекомендацій і служб знайомств

Ключові слова: контент, інформаційний ресурс, контент-аналіз, лінгвістичний аналіз, морфологічний аналіз, соціальна мережа

Предложен подход к разработке системы понимания личности через контент-анализ информационных ресурсов. Использована модель Big-Five на основе пользовательской поведения в социальных сетях. Для определения психологических диспозиций разработан метод анализа англоязычных и украиноязычных постов. Система предназначена для рекрутинга, маркетинга, социальных сетей и Web-сервисов. Анализ аудитории улучшает эффективность контекстной рекламы, систем рекомендаций и служб знакомств

Ключевые слова: контент, информационный ресурс, контент-анализ, лингвистический анализ, морфологический анализ, социальная сеть

UDC 004.89
DOI: 10.15587/1729-4061.2016.77174

THE METHOD OF FORMATION OF THE STATUS OF PERSONALITY UNDERSTANDING BASED ON THE CONTENT ANALYSIS

V. Lytvyn
Doctor of Technical Sciences, Professor*
E-mail: yevhen.v.burov@lpnu.ua
P. Pukach
Doctor of Technical Sciences, Associate Professor**
E-mail: petro.y.pukach@lpnu.ua
I. Bobyk
PhD, Associate Professor**
E-mail: igor.bobyk@gmail.com
V. Vysotska
PhD, Associate Professor*
E-mail: victoria.a.vysotska@lpnu.ua
*Department of Information Systems and Networks***
Department of Mathematics*
***Lviv Polytechnic National University
S. Bandery str., 12, Lviv, Ukraine, 79013

1. Introduction

One of the tasks of the research into content created by the Internet users (posts at the forums, comments of events, profiles in social networks, etc.) is determining their psychological state. Potential customers of such studies are recruiting and marketing companies. Collected and analyzed information about users is used when hiring or promoting products/services.

Automated compilation of personality models of the users is helpful for social networks and Web services. It improves the quality and efficiency of context advertising, referral systems and dating services [1]. In-depth knowledge of the audience is crucial for business and recruiting [2]. Hence, the task of development of information systems of processing the Internet content with the purpose of categorizing users according to certain characteristics. Based on these characteristics, we obtain the user's psychological state.

We propose to choose the "Big Five" of such characteristics (dispositions) [3]: extraversion/introversion (orientation of a person on the outside world, talkativeness, sociability or immersion into the world of imagination and reflection); amiability (the ability to mutual assistance, mutual collaboration and sympathy in relation to others); integrity (discipline, diligence and focus on result); neuroticism (degree of emotional stability, level of control of impulses and anxiety); openness to experience (degree of intellectual curiosity, the desire for new and diverse experiences, impressions). Each personality is ranked by five dispositions [4]. Thus psychologists make up a personality model. Such a model is used to predict the actions of a human, formation of conclusions about his/her professional suitability, prospects for professional growth, opportunities to work in a team, etc. [5].

When recruiting, a number of organizations perform mandatory psychological testing of candidates for compiling their personality model [6]. This is a fairly long procedure that requires certain resources and time costs. Most often, in addition to the interview, they use the users' profiles in social networks for analysis [7]. Automated processing of user-generated content considerably simplifies the process of recruiting [7].

Similarly, for a marketing company to manually analyze potential audience of consumers of advertised products is quite a costly process [8]. Automated analysis of audience through thematic activity in social networks significantly facilitates the process of promotion of products to the market [9]. It also accelerates expanding target audience of consum-

ers of advertised products [10]. Serious drawback of automation of such a process is the lack of automated linguistic analysis of texts and the lack of appropriate dictionaries of stop-words to identify a personality model of the Internet users [11]. The process of content analysis when determining the gender or the age of potential consumer is complicated by his/her posts.

2. Literature review and problem statement

Paper [3] describes the ways to identify each of the 5 personality dispositions by analyzing his/her activity in a social network [4]. Each disposition is determined based on the numerical parameters of person's activity in a social network [5]. For example, the level of integrity (discipline, diligence, and focus on result) is determined by the number of posts with questions, requests for aid [11]. The sign of extraversion is the large number of emoticons (emograms, or the most commonly used in everyday life is the term smile/smiley – a schematic representation of a human face that is used to express emotions) [12]. The frequency of status updates indicates openness to experience [13] and the number of posts that have caused negative assessment from others determines the level of neuroticism [12]. Paper [13] describes in more detail general concepts of personality analysis through his/her activity and profile in social networks [9] as well as the activity of the community [10]. Based on these studies, we developed the method and software for automated determining of psychological disposition of personality. In general, to form the status of psychological state of a personality based on content analysis, it is necessary to address the following 4 tasks:

1. Collect content from various sources from the Internet.

2. Process the content at initial level. Remove the tags, function words, signs, special symbols, hyperlinks, pictures, etc. from the text. Sort out the content (comments to the comments, likes, posts) according to statistics over a specific period of time. Content filtering with spam identification, detection of duplication, content formatting, etc.

3. To conduct content analysis.

4. To perform classification by stop-words (markers).

Theoretical research into content processing is associated with IT development of the analysis and integration of structured, poorly structured and unstructured text data sets of thematic orientation. Thus, for the content collection they use the methods and algorithms of parsing of information sources [14]. Article [15] explored and developed mathematical models for processing and integration of electronic information flows. To process the content, the methods of mathematical statistics, content monitoring are used, while Internet marketing methods are applied for the analysis of conversion of information resources; SEO technology, models of lifecycle content, the Porter stemming and content analysis to process text data arrays, etc. [16]. For the analysis of texts, the Zipf law is used – an empirical regularity of distribution of frequency of words of natural language in texts for its analysis [15]. The corporations EMC, IBM, Microsoft Alfresco, Open Text, Oracle and SAP have developed specifications of Content Management Interoperability Services at the interface of Web-services, to provide for the interaction between the systems of processing the content of information resources. From a scientific standpoint, this segment of IT is insufficiently explored. Each particular project is realized practically from scratch, actually based on own ideas and solutions. The literature covers, in a limited way, theoretical substantiation, research, conclusions, recommendations, generalization for IT development of processing and integration of the content from various information resources. There is an urgent necessity to analyze, summarize and justify existing approaches to realization of such IT. Relevant is the task of creating a set of technological means based on theoretical substantiation of methods, models and principles of processing and integration of the content from different information resources, based on the principle of open systems, which make it possible to manage the process of improving the accuracy of analysis of large volumes of text content.

We will pay more attention to the fourth task, since the correct classification of users by their psychological traits directly depends on it. For this purpose, we first construct a formal model of the system of formation of the status of psychological state of a personality based on the content analysis. Then we will describe the classification method by stop-words (markers). In conclusion, we will verify the proposed method in practice.

3. The purpose and objectives of the study

The aim of this work is to develop a method of automated determination of psychological disposition of personality.

To achieve the goal, the following tasks were formulated:

1. To develop a formal model of formation of the status of psychological state of a personality based on content analysis.

2. To develop a formal model of classification by stop-words (markers) of the content of a social networks' user.

3. To develop the rules of introduction of terms chain of the English and Ukrainian languages for the content analysis of relevant users' posts.

4. To develop the software for automated determination of psychological disposition of personality to conduct experimental research and analysis of the research results of the proposed approach of formation the status of psychological state of personality based on the content analysis.

4. A formal model of the system of formation of the status of psychological state of a personality based on content analysis

Thus, the purpose is to determine the optimal method of automated processing of a set of the Ukrainian/English text content to identify meaningful key words among the existing marked words for automated classification of psychological state of the author of this content. This process is based on the use of the analysis of syntax/semantics of text through content analysis taking into account availability of the marked words. The processing of the set of content to identify meaningful keywords is based on the principle of finding keywords by content (terms), based on the Zipf law and comes down to selecting the words with an average frequency of occurrence.

We will present the system model S of the formation of the status of psychological state of a personality based on the content analysis based on the text data sets of this personality (for example, comments in social networks) by the tuple

$$S =< X, Ident, C, ContProc, Q, Const, PrCont, PersPref, AutAd, ContIntegr, Y >, \quad (1)$$

where x is the incoming data from personalities of social networks, the psychological state of whom is analyzed (history, profile, posts, comments, likes, community, etc.), Ident is the process of identification of the system users and personalization of personalities, C is the content of the system, ContProc is the process of initial processing of the content (content and spam filtering, spam identification, analysis, saving, elimination of duplication, content blocking, etc.), Q are the requests from users, Const is the process of provision of consistency of the content, PrCont is the provision of analysis of private content, PersPref is the analysis of personal preferences and personal data of the user, AutAd is the provision of analysis of automatic settings and user profile updates, ContIntegr is the provision of integration of data from other systems, including those from other social networks, Y are the results of the users' queries concerning the status of psychological state of a personality.

The process of generating answer to the user S in the form of the status of psychological state of the analyzed personality by the main characteristics of the big five is described by superposition of the main functions (input data of one function are the original data of another one) from (1) as follows

$$Y = ContProc \circ PersPref \circ Const \circ \\ \circ AutAd \circ ContIntegr \circ Ident \quad (2)$$

in this case, the main process is ContProc, which is described by the formula

$$Y = ContProc(X,Q,C) = ContAnal \circ ContSav \circ \\ \circ ContBlock \circ ContDupl \circ ContSpFilt \circ SpIdent, \quad (3)$$

de ContAnal is the content analysis, input data/requests, ContSav is the saving of content/results, ContBlock is the content blocking, ContDupl is the elimination of duplication, ContSpFilt is the filtration of content/spam, SpIdent is the identification of content/spam. The process of ensuring privacy of the content PrCont is described by superposition

$$C = PrCont(X,Q,C^{Pc},C^{Pl}) = \\ = ElectrTranst \circ ContSear \circ ContAccs \circ Secrt, \quad (4)$$

where C^{Pc} is the public content; C^{Pl} is the personal content, ElectrTranst is the transactionality; ContSear is the provision of search capabilities; ContAccs is the access to data, Secrt is the provision of security of personal data and conducted transactions.

The level of detail and transactions control Secrt differs depending on the social network, but their required settings are

$$C^{Pl} = Secrt(X,Q,C^{Pc},C^{Us}) = \\ = ContPriv \circ ContLim \circ ContAvail, \quad (5)$$

where C^{Us} is the content of the user, ContPriv is the user privacy provision at the request of the user; ContAvail is the provision of accessibility properties for the content, i.e., even unregistered users can see it; ContLim is the limit for the visibility of content for:

a) the people who are on the contacts list;

b) for specific groups of users of the service;

c) only for the users subscribed to the service.

The process ContIntegr, provision of integration with data from other systems, including those from other social networks, is implemented by appropriate methods and described by superposition

$$C^{Us} = ContIntegr(X,Q,C^{Pc},C^{Pl}) = \\ = OthCollab \circ ContAdThPart \circ PresDevel \circ \\ \circ ContctSup \circ MessSend \circ ContViSear \circ \\ \circ PtofDisp \circ PtofForm \circ ContDownl, \quad (6)$$

where ContctSup is the supporting and establishing of social and friendly contacts with other people; PresDevel is the provision of self-positioning in the network, creation and promotion of online presence by more contacts; ContViSear is the way the content is viewed and searched for; PtofDisp is the way to represent online profile; PtofForm is the author's profile formation; ContDownl is the uploading of own content; ContAdThPart is the addition and sharing of the content by a third party; MessSend is the provision of option to submit public and private messages; OthCollab is the collaboration with other people through social networks.

The users make use of the contacts list for different purposes for the efficient and quality cooperation with other people through social networks through the process of OthCollab. That is why the process to enable support and establishment of social and friendly contacts with other people ContctSup will be presented by superposition

$$C^{Pf} = ContctSup(X,Q,C^{Pc},C^{Pl}) = ContctRepr \circ \\ \circ ContctSear \circ ProfRecom \circ VacanPubl \circ \\ \circ GroupCreat \circ ProfResum, \quad (7)$$

where C^{Pf} is the content of user's profile; ContctRepr is the process of submission through existing contacts and the ability to extend ties; ContctSear is the process to search for companies, individuals, groups of interest; ProfResum is the process of submitting a professional resume and search for job/collaboration; ProfRecom is the process of recommendation and of being recommended; VacanPubl is the process of inviting to groups; GroupCreat is the process of creating a group of interests.

5. Classification by stop-words (markers)

We suggest using the algorithms of analysis of the syntax of the Ukrainian and English-language text for processing and content analysis (stage 3 of the algorithm) of large arrays of text data for finding and analyzing the marked words. We will focus on the features of this particular social network as the source of data for analysis and determination of personality dispositions. With this purpose we will present the main processes of the S system as PersPref, AutAd, Ident and will detail them by superposition

$$C^{St} = PersPref \circ Const \circ AutAd \circ ContIntegr \circ Ident, \quad (8)$$

where C^{St} is the content as a result of statistical data of the activity of a personality.

We will present the process of identification of users Ident as superposition

$$C^{Sp} = Ident(X,Q) = GamCrt \circ \\ \circ MessSeRe \circ UsAuth \circ UsReg, \quad (9)$$

where C^{Sp} is the content of the community of personality and his/her reaction to this content; UsReg is the registration process of users; UsAuth is the user authorization process; MessSeRe is the process of analysis of posts/comments; GamCrt is the process of creating a psychological profile of the personality.

The process of AutAd of provision of automatic settings and user profile updates will be presented by superposition

$$C^{Pf} = AutAd(X,Q,C^{Sp},C^{US}) = \\ = ProfSav \circ GamEd \circ GamCont \circ ProfEd, \quad (10)$$

where ProfEd is the user profile editing, ProfSav is the saving of user profile, GamCont is the entry of new data about the personality according to the analysis of his/her comments, GamEd is the editing of the source list of text content of the analyzed personality.

The process PersPref of the analysis of personal preferences and personal data of the user will be represented by superposition

$$C^5 = PersPref(X,Q,C^{Pf},C^{Pl},C^{Sp},C^{US},C^{Pc},C^{Mr}) = \\ = MatchPred \circ GamPred \circ ProfProc \circ \\ \circ SitChan \circ GamModer \circ SitAdm, \quad (11)$$

where $C^{Mr} \subseteq C$ is the set of marked words in the content of analyzed personality, ProfProc processing the user profile and the profiles of participants of the experiment, SitChan is the editing of dictionaries, GamModer is the moderation of the rules of content monitoring of text data arrays of a specific individual, content analysis to find the marked words, analysis of the text's syntax and semantics, as well as the rules of formation of the status of psychological state of a personality, SitAdm is the system administration, GamPred is the obtaining of result of formation of the status of psychological state of a personality based on the associative rules, MatchPred is the formation of the status of psychological state of a personality based on associative rules [17].

The list of marked content C^{Mr} at $C^{Mr} \subseteq C$ (high frequency of occurrence of the marked words in the comments and posts of the analyzed personality) is a list of marked words that are frequently used (LMWFU) [18]. If $C^{Mr} \subseteq C$ and $P(C) \geq c$, then $P(C^{Mr}) \geq c$. The magnitude L_k defines the set of all lists of marked words in the comments and posts of the analyzed personality from the marked words that are frequently used. The result gives the required set of LMWFU. Associative rule must satisfy the constraint: authenticity

$$(C \Rightarrow C^{Mr}) > c,$$

$$(C \Rightarrow C^{Mr}) = count(C \cup C^{Mr}) / count(C),$$

with $count(C \cup C^{Mr})$ is the number of transactions containing $C \cup C^{Mr}$, and count(C) is the number of transactions containing C. Associative rules are generated for each nonempty LMWFU X, considering all nonempty subsets. Also, for each nonempty subset $C \subset X$ we assign the rule $C \Rightarrow C^{Mr}$, where $C^{Mr}=X \backslash C$ if

$$\frac{cout(X)}{cout(C)} \geq c.$$

The component of the rules of content-monitoring GamModer is the content search and content analysis of the text. The content analysis is aimed at searching for the content in the data set by universal linguistic units. The unit of account is a quantitative measure of the unit of analysis, which allows registering the frequency (regularity) of occurrence of indicator of the category of analysis in the text. Then the text is analyzed for the presence of certain marked words and the results are categorized according to psychological metrics (consciousness, friendliness, extraversion, emotionality and openness to experience) [3], namely

$$C^5 = MatchPred(X,Q,C,P,D,B) = \\ = Opn \circ Cns \circ Ext \circ Agr \circ Nrt \circ Filt, \quad (12)$$

where Filt is the process of filtration of the original text, P is the glossary of rules, D are the dictionaries for classification of the text by psychological dispositions of a personality, B is the dictionary of blocked words, C^5 is the result of analysis of text arrays data and construction of the "Big Five" model, i. e. the hierarchical model of a personality by the five features. In particular, such features are *the openness to experience* $C^{Opn}=Opn(C^{Filt}, U^{Opn}, P, D)$ through parameters U^{Op} (u_1^{Opn} is the frequency of occurrence of words associated with benevolence/malevolence, u_2^{Opn} is the frequency of occurrence of words associated with trust/mistrust, u_3^{Opn} is the frequency of occurrence of words associated with warmth/hostility, u_4^{Opn} is the frequency of occurrence of words associated with sincerity/selfishness); *integrity* $C^{Cns}=Cns(C^{Filt}, U^{Cns}, P, D)$ through parameters U^{Cn} (u_1^{Cns} is the spontaneity/deliberation, u_2^{Cns} is the creativity/narrow–mindedness, u_3^{Cns} is the distinction/mediocrity, u_4^{Cns} is the liberality/parochialism); *extraversion* $C^{Ext}=Ext(C^{Filt}, U^{Ext}, P, D)$ through parameters U^{Ext} (u_1^{Ext} is the sociability/unsociability, u_2^{Ext} is the assertiveness/tranquility, u_3^{Ext} is the activity/passivity); *amiability* $C^{Arg}=Arg(C^{Filt}, U^{Arg}, P, D)$ through parameters U^{Arg} (u_1^{Arg} is the orderliness/negligence, u_2^{Arg} is the thoroughness/carelessness, u_3^{Arg} is the unreliability/reliability) and neuroticism $C^{Nrt}=Nrt(C^{Filt}, U^{Nrt}, P, D)$ through parameters U^{Nrt} (^{Nrt} is the relaxation/nervousness, u_2^{Nrt} is the poise/depression, u_3^{Nrt} is the resistance/irritability).

I. The process of compiling a terms chain *in English* where sentences have strictly defined, special word order [19]. For a simple narrative sentence in English, the main syntactic categories are nominal and verbal groups [20] whose grammatical categories are person and number which determine coordination. The main grammatical characteristics of the components of nominal group in the English language are gender, number, case of noun/pronoun, degree of comparison, person, and of the verbal – person, number, tense, type, method, condition [21].

The nominal group Ñ is expressed by the pronoun P_N or has the following structural scheme (in square brackets we marked optional elements and in braces – elements that can be repeated):

$$Ñ = [D][E][B][\{Ã\}]N[Ẽ], \text{ or } Ñ = N^p, \text{ or } Ñ = ÑQÑ, \quad (13)$$

where D is the determinant, E is the preposition, B is the adverb, Q is the conjunction, Ã is the adjective group, A is the adjective, N is the noun, Ẽ is the prepositions group. The determinant D is a grammatical class (not part of speech), which includes the words (articles, possessive and demonstrative pronouns, quantitative adjectives and nouns in possessive case), which define Ñ in terms of certainty, number, etc.

The adjective group $\tilde{A}$ has the following structural scheme:

$$\tilde{A} = [\{\tilde{B}\}]A[\tilde{E}], \quad (14)$$

where B is the adverb, A is the adjective, E is the preposition, $\check{E}$ is the prepositions group.

The prepositions group $\check{E}$ has the following structural scheme:

$$\tilde{E} = E\tilde{N}. \quad (15)$$

The verbal group $\check{R}$ has the following structural scheme:

$$\tilde{R} = R_V[\tilde{B}], \quad (16)$$

where R_V is the verbal expression, $\tilde{B}$ is the adverbial group. The *verbal expression* R_V consists of lexical, auxiliary and modal verbs R:

$$R_V = [R_V^M][R_V^D][R_V^D][R_V^D]R_V^L, \quad (17)$$

where R_V^M is the modal verb, R_V^D is the auxiliary verb, R_V^L is the lexic verb.

The adverbial group $\tilde{B}$ has the following structural scheme:

$$\tilde{B} = \{B\}[\{\tilde{E}\}]. \quad (18)$$

Let us consider generative grammar for modeling the syntax of a sentence in English of the described structural scheme. The alphabet is a noun group, a verbal group and their components (non-term characters), as well as vocabulary of language (respective term symbols). According to the requirements and rules of the English language, the first place in a sentence is taken by a nominal group and after it, the verbal. All possible transformations of term symbols into the non-term ones make up a set of rules. During the determination, they receive countless number of terms chains in English of the corresponding structural scheme, which is why such grammar will be unlimited and, because of its complexity, will not be implemented. To introduce tne context-dependent grammar, let us model the process of building a simple narrative sentence in English with such constraints on the structure of the sentence:

– $\tilde{N}$ is not expressed by pronoun, the determinant-article, simplified by adjective group and without prepositional group;

– $\check{R}$ is expressed only by lexical verb (type – simple, mood – indicative, condition – active), $\check{E}$ is omitted from $\tilde{B}$.

Let us consider a sentence of the following simplified structural scheme.

1. Simplified **nominal group** $\tilde{N}$**:**

$$\tilde{N} = D\{A\}N, \text{ or } \tilde{N} = N^p. \quad (19)$$

Now the main grammatical characteristics of $\tilde{N}$ are number, case of the noun, the degree of comparison, person. A can be connected by conjunction.

2. Simplified **verbal group** $\check{R}$ will have the following structural scheme:

$$\tilde{R} = R_V[B]. \quad (20)$$

Table 1 indicates components of the simplified $\tilde{N}$, $\check{R}$ and their grammatical categories (in brackets is the equivalent in English, after «/» we show the used denotations). Let us consider the grammar G_1=(V, T, S, P) for the analysis of the English texts (Table 2), where V is the alphabet (dictionaries), term symbols T (the found marked words with regard to the part of speech), # is the symbol of the end of a sentence, S is the initial symbol.

Table 1

Denotations of linguistic variables in English

Type	Description
Denotation of grammatical categories of nominal group	
Nominal group/$\tilde{N}$	determinant/D, adjective/A, noun/N, pronoun/$N^{pronoun}$;
Number/NR	Singular/sg, Plural/pl;
Case/CS	Common Case/cc, Possessive Case/pc;
Comparison Degree/CD	Positive Degree/pd, Comparative Degree/cd, Superlative Degree/sd;
Person/PR	First/1, Second/2, Third/3.
Denotation of grammatical categories of verbal group	
Verbal group/$\check{R}$	verb/R, adverb/B;
Number/NR	Singular/sg, Plural/pl;
Person/PR	First/1, Second/2, Third/3;
Time/TM	Present/pr, Past/ps, Future/ft.

Table 2

Rules for generating terms chain in the English language

No.	Group	Rules
I	S selection	$S \to \#\tilde{N}_{NR,PR}\tilde{R}_{NR,PR}\#$
II	Convolution $\tilde{N}$	1) $\tilde{N}_{NR,PR} \to D\tilde{N}_{NR,PR}$; 2) $\tilde{N}_{NR,PR} \to A_{CD}\tilde{N}_{NR,PR}$; 3) $\tilde{N}_{NR,PR} \to \tilde{N}_{NR,PR}E\tilde{N}_{NR,PR}$; 4) $\tilde{N}_{NR,PR} \to N_{NR,CS}$; 5) $\tilde{N}_{NR,PR} \to \tilde{N}_{NR,PR}Q\tilde{N}_{NR,PR}$; 6) $A_{CD} \to A_{CD}QA_{CD}$; 7) $\tilde{N}_{NR,PR} \to N_{NR,CS}^{pronoun}$; 8) $\tilde{N}_{NR,PR} \to \tilde{N}_{NR,PR}E\tilde{N}_{NR',PR'}$; 9) $A_{CD} \to BA_{CD}$; 10) $A_{CD} \to B^{degree}A_{CD}$ 11) $\tilde{N}_{NR,PR} \to E\tilde{N}_{NR,PR}\tilde{N}_{NR,PR}$; 12) $\tilde{N}_{NR,PR} \to N_{NR,CS}EN_{NR',CS}$
III	Convolution $\check{R}$	1) $\tilde{R}_{NR,PR} \to R_{NR,PR,TM}$; 2) $\tilde{R}_{NR,PR} \to R_{NR,PR,TM}B$; 3) $B \to BQB$; 4) $B \to BB$; 5) $\tilde{R}_{NR,PR} \to \tilde{R}_{NR,PR}\tilde{N}_{NR,PR}$; 6) $BB \to B^{degree}B$
IV	Words	Finding marked words with regard to the part of speech

II. The process of compiling a terms chain *in Ukrainian*, to which a free order of words in a sentence is inherent [22], which, however, does not deny existence of a stable order of certain language elements [23]. For a simple complete sentence with a direct order of words, we will consider the structural scheme as fixed; the main syntactic categories of such a sentence are nominal and verbal groups [24]. The unlimited grammar, built on the same base as that in the previous examples, will have no application due to its complexity [25]. To form the context-dependent grammar, we will introduce certain limitations, first of all, for the structure of the sentence. Based on the rules for constructing sentences in the Ukrainian language by direct order of words (for example, adjective is in preposition to noun, elements of nominal group are grouped around the noun, etc.), let us consider the **nominal group** $\tilde{N}$ of the following structural scheme.

$$\tilde{N} = \{AN\} \text{ or } \tilde{N} = N^{p}. \quad (21)$$

The adjective and the noun in $\tilde{N}$ agree among themselves by case, number and gender [21–24], and are also a grammatical category of the pronoun.

Let us consider the **verbal group Ř** of the following structural scheme:

$$\tilde{R} = R\tilde{N} \text{ or } \tilde{R} = \tilde{N}R. \quad (22)$$

Given the grammatical characteristics of verb in the Ukrainian language, coordination between the nominal and the verbal group is executed by number, gender and person (Table 3). Let us consider the grammar G_3= =(V, T, S, P). The set of rules P will be presented in the form of Table 4. Note that here in the rules IV we did not take into account the coordination A with the animated S in the accusative.

Table 3

Denotation of linguistic variables in the Ukrainian language

Type	Description
Denotation of grammatical categories of nominal group	
Nominal group/Ñ	adjective /A, noun/N, pronoun /N^{pron};
Number/NB	singular/sn, plural/pl;
Gender/GD	male/m, female/f, neutral/n;
Case/CS	nominative/nm, genitive/gn, dative/td, accusative/ac, ablative/ab, locative/lc, vocative/vc;
Person/PS	First/1, Second/2, Third/3.
Denotation of grammatical categories of verbal group	
Verbal group/Ř	verb/R, adjective within a nominal group /A, noun/N;
Number/NB	singular/sn, plural/pl;
Gender/GD	male/m, female/f, neutral/n;
Person/PS	First/1, Second/2, Third/3;
Tense/TN	present/pr, past/ps, future/ft.

In thematic dictionaries, next to each word is its property (Table 5), where V is the adjective, A is the verb, and the groups a b c d o describe certain nouns (Fig. 1).

Table 4

Rules of construction a sentence in Ukrainian

No.	Selection	Rules
I	S	$S \to \#\tilde{N}_{GD,NB,nm,PS}\tilde{R}_{NB,pr,PS}\#$
II	Ñ	1) $\tilde{N}_{GD,NB,CS,3} \to \tilde{N}_{GD,NB,CS,3}\tilde{N}_{GD',NB',gn,PS'}$; 2) $\tilde{N}_{GD,NB,CS,3} \to A_{GD,NB,CS}\tilde{N}_{GD,NB,CS,3}$; 3) $K_1\tilde{N}_{GD,NB,CS,PS}K_2 \to K_1\tilde{N}^{pron}_{GD,NB,CS,PS}K_2$ where K_1 is the symbol different from the symbol $A_{GD,NB,CS}$, and K_2 is the symbol different from the symbol with index $CS' = gn$ 4) $\tilde{N}_{GD,NB,CS,3} \to N_{GD,NB,CS}$; 5) $\tilde{N}_{GD,NB,CS,3} \to E\tilde{N}_{GD,NB,CS,3}$; 6) $\tilde{N}_{GD,NB,CS,3} \to \tilde{N}_{GD,NB,CS,3}\tilde{N}_{GD,NB,lc,3}$
III	Ř	1) $\tilde{R}_{NB,pr,PS} \to R_{NB,pr,PS}\tilde{N}_{GD',NB',ac,PS'}\tilde{N}_{GD'',NB'',ab,PS''}$; 2) $\tilde{R}_{NB,pr,PS} \to R_{NB,pr,PS}\tilde{N}_{GD',NB',ab,PS'}\tilde{N}_{GD'',NB'',ac,PS''}$; 3) $\tilde{R}_{NB,pr,PS} \to R_{NB,pr,PS}\tilde{N}_{GD',NB',ac,PS'}$; 4) $\tilde{R}_{NB,pr,PS} \to R_{NB,pr,PS}\tilde{N}_{GD',NB',ab,PS'}$; 5) $\tilde{R}_{NB,pr,PS} \to R_{NB,pr,PS}E\tilde{N}_{GD,NB,lc,3}$; 6) $\tilde{R}_{NB,pr,PS} \to E\tilde{N}_{GD,NB,lc,3}R_{NB,pr,PS}$
IV	words	Finding the marked words with regard to the part of speech

Table 5

Excerpts from thematic dictionary of computer topics

Excerpt 1	Excerpt 2	Excerpt 3
to buffer/ABGH	keyboard/V	a console/ij
to format/AB	COBOL/e	configurator/efg
to decode/ABGH	codec/efg	copyleft/e
to cache/ABGH	coder/efg	copyright/e
cyrillic/V	code-generator/efg	cryptographic/V
kilobite/V	code-compatible/V	crypto protected/V
a kilobite/efg	combolist/ab	cross-assembler/efg
kilobit/V	commuted/V	cross-compiler/efg
a kilobit/efg	concatenation/ab	cookie/ab
kilobaud/efg	console/V	cursor/V

```
Файл  Правка  Вид  Справка
#####################################################################
# Групи a b c d o
#
# -- Перша відміна: іменники жіночого та чоловічого та середнього роду
#
# -- Друга відміна: іменники чоловічого роду із закінченням на -ар -ир
#                   наголошені (Мішана група на -ар -ир)
#
# -- Друга відміна: іменники чоловічого роду з чергуванням -і -о
#
# -- Числівники -ять, -сят, -сто
#
#

SFX a Y 235

#
# ОДНИНА (множина перенесена в гр. b)
#
# Спочатку перша відміна
#
# тверда група в Називному відмінку однини з закінченням на -а
# однина
SFX a   а и [^жчшщ]а  # хата  хати  (Р.)
SFX a   а і [^гґкх]а  # хата  хаті  (Д.М.)
SFX a   а у а    # хата  хату  (З.)
SFX a   а ою [^жчшщ]а  # хата  хатою  (О.)
```

Fig. 1. Dictionary of nouns

Fig. 2, *a* presents a list of rules for reducing a word to the basic from, where the attribute flag defines the type of a word (in the given example, a nominal group, singular), the attribute mask – the rule of identification of the ending of a word, the attribute find – the ending of a word in the nominative case, the attribute repl – the ending of a word at conjugation. In brackets are the exceptions.

id	ordering	state	flag	type	lang	mask	find	repl
26	26	1	a	SFX	uk	ін	ін	оном
27	27	1	a	SFX	uk	ін	ін	оні
28	28	1	a	SFX	uk	ir	ir	огу
29	29	1	a	SFX	uk	ir	ir	огові
30	30	1	a	SFX	uk	ir	ir	огом
31	31	1	a	SFX	uk	ir	ir	озі
32	32	1	a	SFX	uk	[^л]ід	ід	оду
33	33	1	a	SFX	uk	[^л]ід	ід	одові
34	34	1	a	SFX	uk	[^л]ід	ід	одом
35	35	1	a	SFX	uk	[^л]ід	ід	оді
36	36	1	a	SFX	uk	[^пг]лід	ід	ьоду
37	37	1	a	SFX	uk	[^пг]лід	ід	ьодові
38	38	1	a	SFX	uk	[^пг]лід	ід	ьодом
39	39	1	a	SFX	uk	[^пг]лід	ід	ьоді
40	40	1	a	SFX	uk	[пг]лід	ід	оду
41	41	1	a	SFX	uk	[пг]лід	ід	одові
42	42	1	a	SFX	uk	[пг]лід	ід	одом
43	43	1	a	SFX	uk	[пг]лід	ід	оді
44	44	1	a	SFX	uk	іб	іб	обу

a

id	ordering	state	word	lang
1	1	1	після	uk
2	2	1	між	uk
3	3	1	are	en
4	4	1	and	en
5	5	7	між	uk
6	6	1	been	en
7	7	1	has	en
8	8	1	their	en
9	9	1	any	en
10	10	1	the	en
11	11	1	with	en
12	12	1	таких	uk
13	13	1	їхніми	uk
14	14	1	как	ru
15	15	1	такої	uk

b

Fig. 2. List: *a* – the rules of reducing a word to the basic form; *b* – blocked words

For example, the first line describes a particular example

```
# Nouns ending in –ін with alternating -i -o
SFX a ін ону        ін      # загін загону          (Д.Р.)
SFX a ін онові      ін      # загін загонові        (Д.)
SFX a ін оном       ін      # загін загоном         (О.)
SFX a ін оні        ін      # загін загоні          (М.)
the third line describes
# Nouns ending in –ir with alternating -i -o
SFX a ir огу        ir      # батір батогу          (Д.Р.)
SFX a ir огові      ir      # батір батогові        (Д.М.)
SFX a ir огом       ir      # батір батогом         (О.)
SFX a ir озі        ir      # батір батозі          (М.)
the ninth line describes
# Nouns ending in –ід with alternating -i -o
SFX a ід оду        [^л]ід  # провід проводу        (Д.Р.)
SFX a ід одові      [^л]ід  # провід проводові      (Д.)
SFX a ід одом       [^л]ід  # провід проводом       (О.)
SFX a ід оді        [^л]ід  # провід проводі        (М.)
```

Fig. 2, *b* demonstrates an example of dictionary of the words blocked by the moderator, that is, the words that cannot be the keywords, but their consideration in the analysis of texts significantly affects the final result. The search for the required words in the comments of users of social networks is presented in Fig. 3. By using the data from these three tables and based on the analysis, the table "Results of analysis" is compiled.

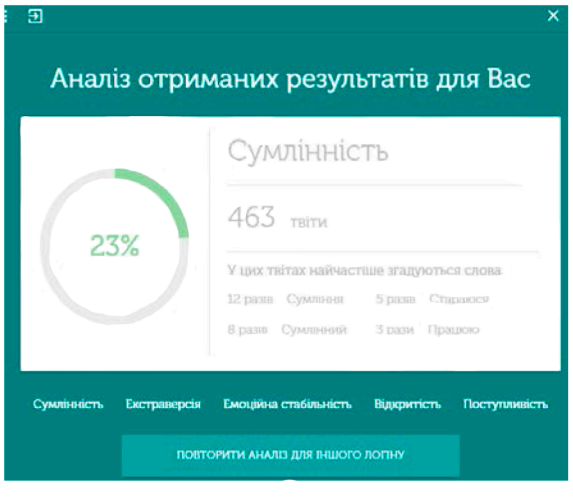
```
function SubstringSearch(sub, str)
{
    var i, j, n = sub.length,
        N = str.length - n + 1;

    for (i = 0; i < N; i++)
    {  j = 0;
       while (j < n && sub.charAt(j) === str.charAt(i+j)) j++;
       if (j === n) return i;
    }
    return -1;
}
```

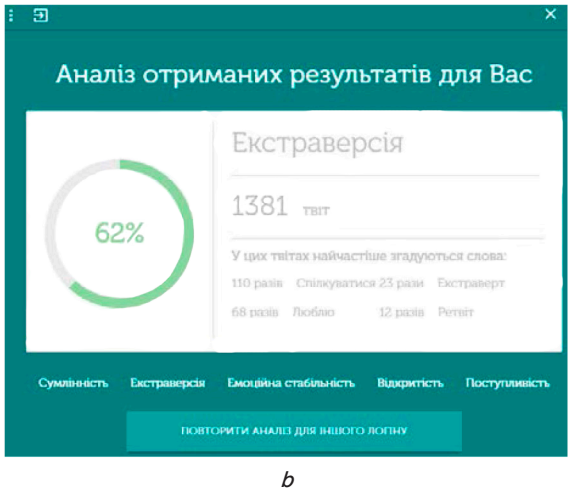
Fig. 3. Algorithm of search for words in comments

7. Results of the research into the proposed approach of formation the status of psychological state of a personality based on the content analysis and their discussion

In the course of the research, we developed the IS definition of psychological analysis of a personality based on the features of the "Big Five". The system operates by analyzing the messages of users in the social network. IS was developed in the form of a desktop program, which is the Internet service at the same time, and allows analyzing psychological state of a particular user of a social network by his/her messages. All collected results are registered in the database. The results are displayed in the form of percent ratio for each trait, the number of tweets, as well as the most frequently used words related to these traits (Fig. 4, *a*, *b*).



a



b

Fig. 4. Main superscription: *a* – integrity; *b* – extraversion

The developed system was tested on 100 students of the 2nd and 3rd year of study at the Department of Information Systems and Networks of the National University "Lvivska Politekhnika" (Lviv, Ukraine) by examining their pages in social networks. In addition, we conducted a regular survey by psychological tests to determine their traits of the "Big Five". For each of the 5 characteristics we computed the correlation coefficient between the data obtained using questionnaire and defined based on the content analysis of the students' pages in social networks (Table 6). If one

considers that the results of the survey are 100 % authentic, then the values obtained by automated content analysis of the pages in social networks are satisfactory by 4 indicators because their correlation coefficient exceeds 0.7 (at the value of the correlation coefficient larger than 0.7, the connection is considered to be strong) [26]. The correlation coefficient of the metric "integrity" was lower than 0.7 and, therefore, such a connection is moderate. From the resulting correlation coefficient, we can conclude that the system produces a satisfactory result. With minor changes, it can be used when searching for employees for certain positions. The developed program complex demonstrates satisfactory results because we clearly defined stop-words for each of the traits of the "Big Five" and the rules of linguistic processing of the English and Ukrainian natural languages tests, as well as the rules for classification the user's content in social networks by the stop-words (markers).

Table 6

Coefficients of correlation between the results of the survey and received by the automated system based on the content analysis of pages in social networks

Analysis parameters	extra-version/ introver-sion	ami-ability	integrity	neuroti-cism	Openness to experi-ence
Coefficient of correlation	0,87	0,9	0,64	0,88	0,73

Such an information system is recommended to use for formation the status of psychological state of a personality based on the content analysis. Automated analysis of messages of the users in a social network reduces by almost twice the time of finding a potentially promising employee among the job seekers taking into account his/her psychological portrait for a specific position. Manual search and analysis of the activity of a particular (known in advance) person, even by skilled qualified experts, such as psychoanalysts, is a cumbersome and routine process. And if searching for and analyzing the information about a potential employee manually in his/her comments and posts among a multitude of those applied, then the time cost grows exponentially. Automated search and analysis of information, both messages and posts of users in a social network, significantly filters out the amount of received data, gives a clearer set of exact data without information noise and generates the required content in the form of a psycho-questionnaire of a particular analyzed person. The system cannot completely replace the recruiters, psychologists, analysts and the people conducting interviews. It can only help them in organizing their work, reducing the time to process the data and more precisely collect the necessary information with a specific purpose.

5. Conclusions

1. The model of information system is proposed of determining the psychological state of personalities based on the five personality dispositions (extraversion/introversion, amiability, integrity, neuroticism, openness to experience), which is based on the content analysis of the Internet resources where users leave their mark (social networks, forums, chats, etc.).

2. To determine the psychological dispositions of a personality, we developed and described the method of search and analysis of the marked words for two languages (English and Ukrainian). Its essence is the usage of the rules of relation of the marked words to psychological dispositions in the content analysis of the English-language posts of the users. Usually psychological dispositions are determined through questioning of the analyzed person. A person may not give false information in questionnaires when aware of it being analyzed. The peculiarity of this method that sets it apart from the questionnaires is the automated analysis of the comments of users of social networks over a long period of time. Usually no one is capable of submitting false information for a long time in everyday life.

3. In the course of research, we created the rules of relation of the marked words and psychological dispositions for the content analysis of the English and Ukrainian posts. Their essence is the automatic classification of the found marked words in the user's texts with certain frequency (set by moderators or psychologists). The found marked words are referred to the corresponding disposition by production rules. The frequency of occurrence of the marked words is calculated in each disposition and in each disposition category, respectively. The more the weight of a disposition category is, the more likely this category corresponds to the character of the analyzed person.

4. The information system is created of determining the psychological state of a personality, based on the developed approach and the methods of the content processing. It is, locally available on the server of the Department of Information Systems and Networks of the National University "Lvivska Politekhnika", the system for collecting statistical data and conducting experimental research and analysis of the research results of the proposed approach of formation of the status of psychological state of personality based on the content analysis.

References

1. Lovakov, A. Otsenka lichnosti po aktivnosti v sotsialnyh setyah ili Big Data prihodyat v psihologiyu [Electronic resource] / A. Lovakov. – Available at: http://psyresearchdigest.blogspot.com/2013_11_01_archive.html – Title from the screen.
2. Alizar, A. Sostavlenie modeli lichnosti po aktivnosti v sotsialnoy seti [Electronic resource] / A. Alizar. – Available at: https://xakep.ru/2012/04/26/58618/ –Title from the screen.
3. Bai, S. Big–Five Personality Prediction Based on User Behaviors at Social Network Sites [Electronic resource] / S. Bai, T. Zhu, L. Cheng. – Available at: http://arxiv.org/pdf/1204.4809v1.pdf, http://arxiv.org/abs/1204.4809 – Title from the screen.
4. Bai, S. List of computer science publications by Shuotian Bai [Electronic resource] / S. Bai. – Available at: http://dblp.unitrier.de/pers/hd/b/Bai:Shuotian – Title from the screen.
5. The Personality Insights models [Electronic resource]. – Available at: https://www.ibm.com/watson/developercloud/doc/personality-insights/models.shtml – Title from the screen.

6. Solovyov, D. Potrebnosti i povedenie lyudey v sotsialnyh setyah. Teoriya “laykov” [Electronic resource] / D. Solovyov. – Available at: http://www.cossa.ru/234/13291/ – Title from the screen.
7. Jeffrey, M. (2012). Recruiting 5.0: Psychological profiles on social networks [Electronic resource] / M. Jeffrey. – Available at: http://www.eremedia.com/ere/recruitment-5-0-the-future-of-recruiting-the-final-chapter/ – Title from the screen.
8. Prokhorov, A. Sotsialnye seti: psihologiya, sotsiologiya, biznes [Electronic resource] / A. Prokhorov. – Available at: http://compress.ru/article.aspx?id=23890 – Title from the screen.
9. Global Web Index. Social Web Involvement [Electronic resource]. – Global Web Index. – Available at: https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnx0b21ob2dlcnN8Z3g6NzVlNjM3MWExNjk3NzAwNA – Title from the screen.
10. Bennett, J. Visualization Critique [Electronic resource] / J. Bennett. – Available at: http://vizthinker.com/visualization-critique/ – Title from the screen.
11. Kluemper, D. H. Social Networking Websites, Personality Ratings, and the Organizational Context: More Than Meets the Eye? [Text] / D. H. Kluemper, P. A. Rosen, K. W. Mossholder // Journal of Applied Social Psychology. – 2012 – Vol. 42, Issue 5. – P. 1143–1172. doi: 10.1111/j.1559-1816.2011.00881.x
12. Schwartz, H. A. Personality, Gender, and Age in the Language of Social Media: The Open–Vocabulary Approach [Text] / H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal et. al. // Plos One. – 2013 – Vol. 8, Issue 9. – P. e73791. doi: 10.1371/journal.pone.0073791
13. Kosinski, M. Private traits and attributes are predictable from digital records of human behavior [Text] / M. Kosinski, D. Stillwell, T. Graepel // Proceedings of the National Academy of Sciences. – 2013 – Vol. 110, Issue 15. – P. 5802–5805. doi: 10.1073/pnas.1218772110
14. Lande, D. Osnovy integratsii informatsionnyh potokov [Text]: monograph / D. Lande. – Kyiv: Inzhiniring, 2006. – 240 p.
15. Lande, D. Osnovy modelirovaniya i otsenki elektronnyh informatsionnyh potokov [Text] / D. Lande, V. Furashev, S. Braichevsky, O. Grigorev. – Kyiv: Inzhiniring, 2006. – 348 p.
16. Clifton, B. Google Analytics: professional analysis of attendance of web sites [Text] / B. Clifton. – Moscow: OOO «ID Williams», 2009. – 400 p.
17. Shahidi, A. Introduction to the analysis of association rules [Electronic resource] / A. Shahidi. – Available at: https://basegroup.ru/community/articles/intro – Title from the screen.
18. Association rules search in Data Mining [Electronic resource]. – Available at: http://ami.nstu.ru/~vms/lecture/data_mining/rules.htm – Title from the screen.
19. Beh, P. Anhliyska mova. Samovchytel [Text] / P. Beh, L. Byrkun. – Kyiv: «Lybid», 1993. – 232 p.
20. English Grammar in an accessible narrative [Electronic resource]. – Available at: http://realenglish.ru/crash/lesson3.htm – Title from the screen.
21. English Verbs (Part 1) – Basic Terms [Electronic resource]. – Available at: https://sites.google.com/site/englishgrammarguide/Home/english-verbs--part-1----basic-terms – Title from the screen.
22. Bagmut, A. Poryadok sliv [Text] / A. Bagmut. – Kyiv: M. P. Bazhana «Ukr. Encyclopedia», 2007. – P. 675–676.
23. Zubkov, M. Ukrayinska mova: Universalnyy dovidnyk [Text] / M. Zubkov. – Kyiv: Publishing House «Shkola», 2004. – 496 p.
24. Ukrayinskyy pravopys [Text]. – O. O. Potebnia Linguistics Institute of Ukraine NAS, Ukrainian Institute of Ukraine NAS. Kyiv: Nauk. dumka, 2007. – 288 p.
25. Shulzhuk, K. Syntaksys ukrayinskoyi movy [Text] / K. Shulzhuk. – Kyiv: Academy, 2004. – 397 p.
26. Uilks, S. Matematicheskaya statistika [Text] / S. Uilks. – Moscow: Nauka, 1967. – 632 p.