International Journal of Mathematics and Computational Science
Vol. 1, No. 4, 2015, pp. 174-182
http://www.aiscience.org/journal/ijmcs

Statistical Methods of Virtual Community Users Age Verification

Solomia Fedushko[1, *], Halyna Biluschak[2], Yuriy Syerov[1]

[1]Institute of Humanities and Social Sciences, Social Communications and Information Activities Department, Lviv Polytechnic National University, Lviv, Ukraine
[2]Institute of Applied Mathematics and Fundamental Sciences, Higher Mathematics Department, Lviv Polytechnic National University, Lviv, Ukraine

Abstract

In the article the statistical methods of web-users age verification are investigated and developed by mean of computer-linguistic analysis of virtual community user's information tracks. The issue of personal data verification of virtual community user's accounts is important to web-community administrators. The aim of this method is substantially effects on the efficiency of virtual communities functioning and improving the registration and moderating processes in the virtual community. The set of age indicative features of virtual community users is formed for the age verification of web-users. The age indicative characteristics based on the socio-demographic markers set up by experts. Statistical methods in the learning sample of virtual community users of two Ukrainian web-forums *(Lviv. Forum Ridne Misto* and *Rock.Lviv.Ua)* are presented. The computer-linguistic method of socio-demographic characteristics age-validation in social communications is developed.

Keywords

Socio-demographic Marker, Account, Statistical Method, Virtual community, Data Verification

Received: April 9, 2015 / Accepted: April 24, 2015 / Published online: May 27, 2015

1. Introduction

Virtual communities accumulated a huge database of contacts and profiles, which contain a lot of information about the person. The auditorium of virtual communities is a large number of people regardless of age, gender, occupation, education, ethnicity, social status etc., who in order to register need to fill in a form with their personal data. Nowadays, the necessity of verifying of virtual community users' personal data, especially age of the virtual community user, is the topical issue. The value of this research lies in verifying basic socio-demographic characteristics of communities' user [1, 2] based on statistical analysis of information track. Every socio-demographic characteristic of virtual community user is determined by analysis of linguistic features in virtual community user's communication. The difference in style of writing posts by virtual community users is the basis for developing effective methods of verifying of personal data in user account. Validation of socio-demographic characteristics of online community users is one of the main tasks of developing method of improving the functioning and moderation of virtual communities. In order to avoid of conflicts in the virtual community moderators and administrators need to clearly verify the authenticity of the virtual community users belonging to certain socio-demographic groups. The primary task of this work is to develop statistical methods for verifying the age of virtual community user.

2. Research Significance

Socio-demographic characteristic of the "age" is selected for the age validation specified virtual community users considering important factors:

* Corresponding author
E-mail address: felomia@gmail.com (S. Fedushko)

Real online-threats to internet users aged from 6 to 17 years (disclosure of confidential personal information, access to content that is not responding the age of users and has a negative effect to physical and psychological health of children, online abuse, internet-marketing crimes, etc. [3]);

Elimination of adolescents (adolescents filed an application or had already become a virtual community user that assigned only for adult users and avoiding the registration of adult user in the children virtual community).

For age differentiation of virtual community users a set of age-indicative features (23 features) of virtual community users is formed by the experts.

The age-indicative features formed based on researches, science theories and ideologies of the leading scientists: Arinze B., Ridings C., Gefen D. [4], Schiano D., Chen C., Ginsberg J., Gretarsdottir U., Huddleston M., Isaacs E. [5], Subramanyam K. [6], Rubio M., Berg-Weger M., Tebb S., Lee E.& Rauch S. [7], Bodnar R. [8, 9], Huffaker D. [10], Herring S. [11], Calvert S., Mahler B., Zehnder S., Jenkins A., Lee M. [12, 13], Crystal D.[14], Damer B. [15], Witmer D. [16], Wolf A. [17], Dzyubyshyna-Melnyk N. [18], Aksak V. [19], analysis of web-forums content (*Lviv. Forum Ridne Misto* [20] and *Rock. Lviv. Ua* [21]).

3. Experimental Investigation

3.1. Model of Age-Indicative Features

The classification of determined age-indicative features of virtual community users is follow in Table 1.

Table 1. Classification of age-indicative features of virtual community user

LKIndicator	Indicative features
AGE–A Assertiveness and self-actualization style	AGE–A(1.1) familiar vocabulary
	AGE–A(1.2) vulgarisms
AGE–B Slang variation	AGE-B (1.1) olbanian slang
	AGE-B (1.2) computer slang and Internet slang
	AGE-B (1.3) teen slang
	AGE-B (1.4) attributes of young fashion
AGE–C Modulation of voice and sound similarity	AGE-C (1.1) uppercase character (caps lock), combination case
	AGE-C (1.2) designation of interjections
	AGE-C (1.3) words replacement based on sound similarity
AGE–D Text economy	AGE-D (1.1) truncation
	AGE-D (1.2) acronyms
	AGE-D (1.3) transliteration
	AGE-D (1.4) abbreviation
	AGE-D (1.5) word formation
AGE-E Uncodified units and non-verbal means	AGE-E (1.1) combinations of symbols and letters
	AGE-E (1.2) excessive amounts of punctuation marks and special symbols
	AGE-E (1.3) replacing of letters to figures
	AGE-E (1.4) replacing of letters to non-alphabetic signs
	AGE-E (1.5) combinations of vowel letters
	AGE-E (1.6) compilation of letters
	AGE-E (1.7) sequence of parentheses ")"
AGE-F Deformalization	AGE-F (1.1) graphical smilies
	AGE-F (1.2) familiar personal names

These notations of linguistic-communicative indicators and indicative features we will use in this work.

The set of linguistic-communicative age-indicators of virtual community user for the convenience of analysis is described as follows:

$$\mathrm{LCI}(\mathrm{IO}_i) = \left(\mathrm{LCI}_j(\mathrm{IO}_i)\right)_{j=1}^{N_i^{\mathrm{LCI(IO)}}} \quad (1)$$

where $\left(\mathrm{LCI}_j(\mathrm{IO}_i)\right)_{j=1}^{N_i^{\mathrm{LCI(IO)}}}$ is age-indicative features set of linguistic-communicative indicators of web-users;

$N_i^{\mathrm{LCI(IO)}}$ is number of these age-indicative features of particular linguistic-communicative indicator virtual community user;

$1 \le i \le 6$, as "age" socio-demographic characteristics define six linguistic and communicative indicators. For all investigated socio-demographic characteristics i belongs to the set of integers ($i \in \mathrm{N}$).

According to the established hierarchy the vector of markers indicative feature is defined. Indicative feature determine linguistic-communicative indicator of virtual community users age.

The marker is linguistic and graphic feature which include information about socio-demographic belonging of anonymous virtual community user and identify authenticated web-user or group of users.

This dependence is given by Eq.2:

$$IO = \left(Marker_j\left(IO_i\right)\right)_{j=1}^{N_i^{Marker}} \quad (2)$$

The indicative features of socio-demographic characteristics - "age" of virtual community users according to Eq.1 and Eq. 2 is as follows (see Fig.1):

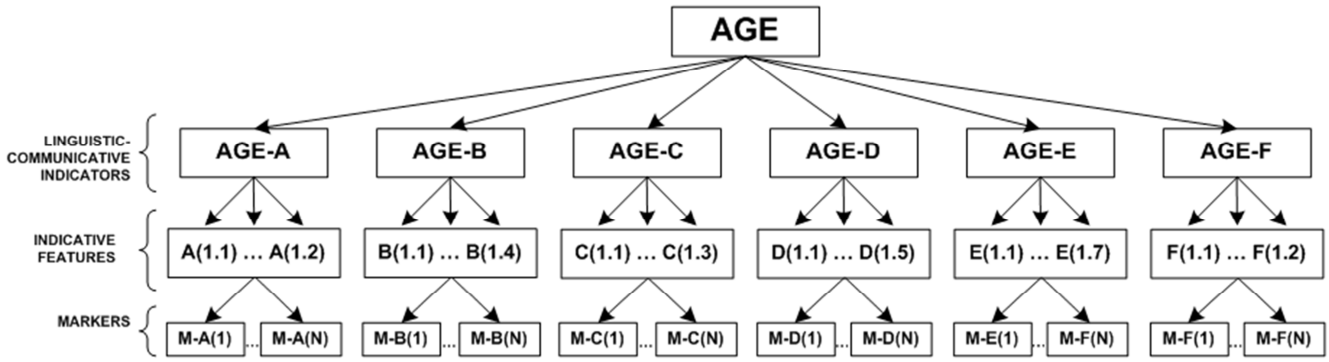

Figure 1. Scheme of age-indicative features classification

3.2. Forming and Processing of the Training Sample of Age-Indicative Features

For investigated methods approbation from set of two web-forums users *Lviv. Forum Ridne Misto* and *Rock.Lviv.Ua* the training sample of 80 most active users of this virtual community is formed. All active users evenly into two groups according to age (40 web-users in each group) are divided.

In conformity with mental human development at age stages in developmental psychology [22], virtual community users into two groups are distributed: "teenager" (from 6 to 17 years old) and "adult" (18+ years old). The "age" socio-demographic characteristics of virtual community's user age takes one of two values: $age \in \{"adult";"adolescent"\}$.

Each user of the training sample is thoroughly chosen to take into account the completeness and authenticity of the information track of web-forum user.

The context of messages and topics of discussion is significantly affected by the results of this study. In view of this fact, the basis of this study is diversified sample of users' messages from all web-forums topics: *Lviv. Forum Ridne Misto* and *Rock.Lviv.Ua*. The all discussions on web-forum threads according to interests and hobbies of adults and adolescents are considered.

3.3. Statistical Analysis of the Training Sample of Age-Indicative Features

The ensuring the minimal recognition error and reliable identification of virtual community user age is lying in the calculation of basic statistical characteristics of training sample.

The complex of mathematical statistics methods for resolving the issue of age verification of users of virtual communities is applied. This complex helps to classify the web-forums users depending on the age category, namely, discriminant analysis, cluster analysis and factor analysis [23]. These calculations using the package of applied programs for statistical data analysis «STATISTICA» [24] and the Statistical Package for the Social Sciences «SPSS» [25] are automated realized.

Results of a comparison of the main numerical characteristics considered indicative features in these two groups indicate the presence of statistically significant differences in most of these parameters ($p > 005$).

The absence of significant differences observed only by means of characteristics such as AGE-F (1.1), AGE-E (1.3), AGE-D (1.1), AGE-D (1.4) AGE-E (1.6). This may explain that some indicative features are poorly expressed, but it is an important differentiating factor in determining of age group of specific web-community user.

According to the results of discriminant analysis these indicative features are not the most informative, because both samples are merged into one and conducted similar studies.

For all factors selected sample is divided into two clusters: Claster 1 (41 adult web-forum users) and Cluster 2 (39 adolescent web-forum users). Since we know the age of web-forum users, the incorrectness clustering analysis is conducted.

The result of this clustering: only one web-forum user that

the Internet presents itself as teenager, but he assigned to Cluster 1 (adult group). This web-user indicated in the analysis as C5, its age of 15 years.

According to virtual communities administrating and moderating are the following scenarios of development of this atomic situation:

– user incorrectly enter age in the account to purposefully infiltrate in teenager virtual community with hidden intentions;

– user accidentally wrong enters his age;

– online-communication style of virtual community users is not appropriate to the web-user age.

Graphically, the clustering procedure in tree diagram is shown (see Figure 2).

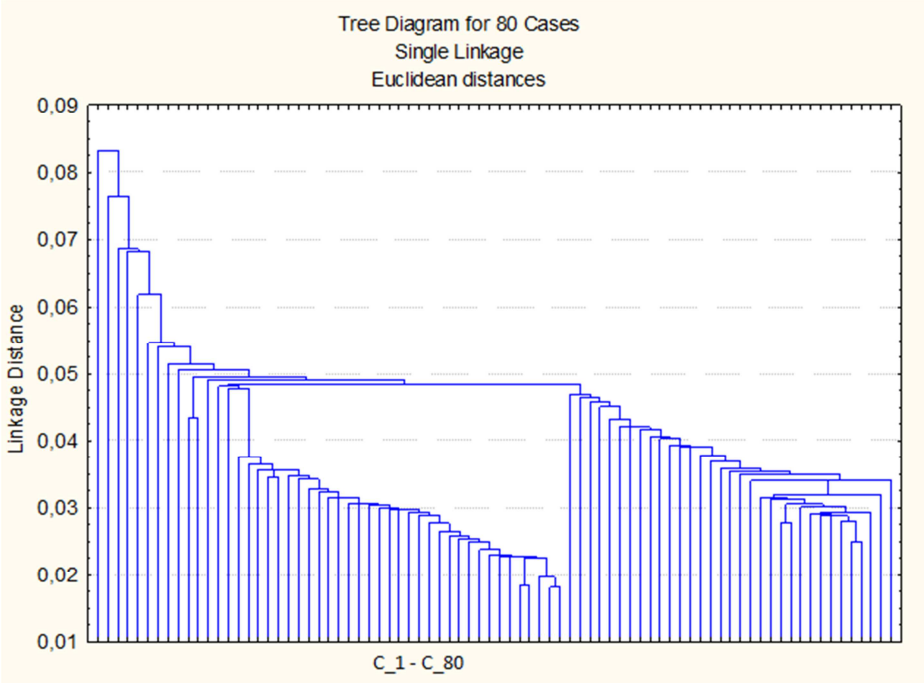

Figure 2. Tree diagram of virtual community users' clasterisation by age

According to the results of the discriminant analysis (see Table 2):

- value of the Wilks' lambda statistic is 0,05255 (is close to zero)
- value of Fisher's F-criterion (23,56) =43, 901 (p<0,0000)

It is show that the discrimination is successful. Classification is correctly conducted.

Also, the classification matrix shows that all objects are classified correctly.

The value of tolerance allows analyzing the informativity of variables (indicative features) from the model.

The value of tolerance indicating a high informative of all indicative features with statistically significant (p<0, 01) are the following indicative features: AGE-A (1.1), AGE-D (1.3), AGE-B (1.1), AGE-E (1.4), AGE-B (1.3).

Table 2. The Results of Discriminant Function Analysis

Discriminant Function Analysis Summary (Sheet1 in Age(80)) No. of vars in model: 23; Grouping: Age (2 groups) Wilks's; Lambda: ,05255 approx. F (23,56)=43,901 p<0,0000							**Standardized Coefficients (Sheet1 in Age(80)) for Canonical Variables**	
N=80	**Wilk's; Lambda**	**Partial**	**F-remove**	**p-level**	**Toler.**	**1-Toler.**	**N=80 (AGE)**	**Coefficient**
AGE–A(1.1)	0,05844	0,924966	4,54277	0,037461	0,653631	0,346369	A(1.1)	-0,34836
AGE–C(1.2)	0,05426	0,996221	0,21241	0,646671	0,765050	0,234950	D(1.5)	-0,18336
AGE–E(1.1)	0,05484	0,985783	0,80762	0,372672	0,711492	0,288508	E(1.3)	-0,14777
AGE–E(1.2)	0,05717	0,945476	3,22941	0,077720	0,597717	0,402283	D(1.1)	-0,09211
AGE–F(1.1)	0,05562	0,971967	1,61512	0,209028	0,715454	0,284546	D(1.2)	-0,07077
AGE–C(1.1)	0,05406	0,999870	0,00730	0,932201	0,620251	0,379749	D(1.4)	-0,06176
AGE–D(1.3)	0,06052	0,893231	6,69378	0,012299	0,700182	0,299818	E(1.6)	-0,0221

Discriminant Function Analysis Summary (Sheet1 in Age(80)) No. of vars in model: 23; Grouping: Age (2 groups) Wilks's; Lambda: ,05255 approx. F (23,56)=43,901 p<0,0000							Standardized Coefficients (Sheet1 in Age(80)) for Canonical Variables	
N=80	**Wilk's; Lambda**	**Partial**	**F-remove**	**p-level**	**Toler.**	**1-Toler.**	**N=80 (AGE)**	**Coefficient**
AGE–E(1.3)	0,05472	0,987802	0,69152	0,409178	0,590552	0,409448	C(1.1)	-0,0149
AGE–B(1.1)	0,07101	0,761295	17,55890	0,000100	0,626359	0,373641	B(1.2)	0,01453
AGE–D(1.1)	0,05428	0,995831	0,23444	0,630136	0,519476	0,480524	E(1.5)	0,02619
AGE–D(1.2)	0,05425	0,996514	0,19593	0,659735	0,736000	0,264000	B(1.4)	0,02656
AGE–E(1.7)	0,05408	0,999538	0,02587	0,872801	0,553488	0,446512	E(1.7)	0,02970
AGE–D(1.5)	0,05534	0,976834	1,32806	0,254046	0,728431	0,271569	C(1.2)	0,07226
AGE–D(1.4)	0,05418	0,997792	0,12391	0,726155	0,611899	0,388101	F(1.2)	0,11502
AGE–B(1.2)	0,05406	0,999893	0,00597	0,938697	0,533812	0,466188	E(1.1)	0,14534
AGE–E(1.4)	0,06948	0,778031	15,97657	0,000189	0,723483	0,276517	A(1.2)	0,18234
AGE–E(1.5)	0,05408	0,999552	0,02511	0,874670	0,690613	0,309387	C(1.3)	0,20238
AGE–A(1.2)	0,05513	0,980505	1,11341	0,295873	0,619839	0,380161	F(1.1)	0,20352
AGE–E(1.6).	0,05408	0,999635	0,02043	0,886862	0,789756	0,210244	E(1.2)	0,31054
AGE–C(1.3)	0,05534	0,976835	1,32801	0,254055	0,597933	0,402067	D(1.3)	0,40150
AGE–B(1.3)	0,06840	0,790352	14,85452	0,000302	0,752680	0,247320	B(1.3)	0,54263
AGE–F(1.2)	0,05440	0,993673	0,35654	0,552839	0,505556	0,494443	E(1.4)	0,56951
AGE–B(1.4)	0,05408	0,999572	0,02397	0,877512	0,641272	0,358728	B(1.1)	0,63473

This is confirmed by the standardized coefficients which indicate the contribution of indicative features of discriminant function value that is one of the approaches to determine the significance of the variable.

Table 1 shows the most significant contribution to the discriminant function with the following factors:

– AGE–B (1.1) (coefficient 0, 63473);

– AGE–E (1.4) (coefficient 0, 56951);

– AGE–B (1.3) (coefficient 0, 54263);

– AGE–D (1.3) (coefficient 0, 40150);

– AGE–E (1.2) (coefficient 0, 31054);

– AGE–A (1.1) (coefficient -0, 34836).

The values of the coefficients indicate that for division into clusters by age the important statistically significant are the following factors: AGE-B (1.1), AGE-E (1.4), AGE-B (1.3), AGE-D (1.3).

According to the results of discriminant analysis can construct a function of classification that allows any new object is attributed to one of these clusters:

F1 = -17,486 + 368,044 * AGE–A(1.1) + 327,17 * AGE–C(1.2) + 135,082 * AGE–E(1.1) - 52,256 * AGE–E(1.2) + 20,93 * AGE–F(1.1) + 296,48 * AGE–C(1.1) + 70,425 * AGE–D(1.3) + 119,03 * AGE–E(1.3) - 137,61 * AGE–B(1.1) + 36,48 * AGE–D(1.1) + 320,039 * AGE–D(1.2) + 81,454 * AGE–E(1.7) + 66,803 * AGE–D(1.5) - 52,98 * AGE–D(1.4) + 268,795 * AGE–B(1.2) + 437,96 * AGE–E(1.4) + 2,778 * AGE–E(1.5) + 39,44 * AGE–A(1.2) + 72,929 * AGE–E(1.6) - 52,719 * AGE–C(1.3) - 38,358 * AGE–B(1.3) - 76,948 * AGE–F(1.2) - 32,126 * AGE–B(1.4)

F2 = -65,087 - 11,348 * AGE–A(1.1) + 400,899 * AGE–C(1.2) + 357,060 * AGE–E(1.1) + 201,953 * AGE–E(1.2) + 161,94 *AGE–F(1.1) +275,77* AGE–C(1.1) + 518,8 * AGE–D(1.3) + 24,88 * AGE–E(1.3) + 818,89* AGE–B(1.1) - 46,996 * AGE–D(1.1) + 248,8 * AGE–D(1.2) + 120,93 * AGE–E(1.7) - 67,234 * AGE–D(1.5) - 80,651 * AGE–D(1.4) + 282,89 * AGE–B(1.2) + 1628,81 * AGE–E(1.4)+ 37,807 *AGE–E(1.5)+ 272,27 * AGE–A(1.2)+ 52,453 *AGE–E(1.6) + 252,83 * AGE–C(1.3) + 588,94 *AGE–B(1.3) + 68,996 * AGE–F(1.2) + 0,901 * AGE–B(1.4)

The graphic representation of clusters confirms effectiveness of the classification (see Fig. 2).

The matrix of the factor structure of the result also allows us to estimate the contribution some factors in the classification.

According to the results we can conclude that the structure of output data is mainly due to the following indicative features as AGE–B (1.3), AGE–E (1.4), AGE–B (1.4), AGE–D (1.3).

Comparison of the studied indicative features for receiving clusters (see Table 3) indicates on the significant difference between mean values of indicative features as AGE–B (1.3), AGE–D (1.3), AGE–E (1.4), AGE–C (1.3). The group of teens (Cluster 2) is characterized by more height mean value than the adult group (Cluster 1), namely:

AGE–B (1.3): mean value 0,032172 for adolescents vs 0,006376 for adults;

AGE–D (1.3): mean value 0,023031 for adolescents vs 0,004444 for adults;

AGE–E (1.4): mean value 0,019441 for adolescents vs 0,005059 for adults;

AGE–C (1.3: mean value 0,017359 for adolescents vs 0,005524 for adults.

Table 3. The Results of Cluster Analysis

Descriptive Statistics for Cluster (Sheet1 in Age(80))	Cluster1 Cluster contains 41 cases			Cluster2 Cluster contains 39 cases		
	Mean	Standard	Variance	Mean	Standard	Variance
AGE–A(1.1)	0,015095	0,005958	0,000036	0,020795	0,008989	0,000081
AGE–C(1.2)	0,020300	0,008143	0,000066	0,028818	0,008054	0,000065
AGE–E(1.1)	0,006808	0,004582	0,000021	0,017126	0,006164	0,000038
AGE–E(1.2)	0,004590	0,004085	0,000017	0,025228	0,013842	0,000192
AGE–F(1.1)	0,018146	0,015116	0,000228	0,022592	0,007176	0,000051
AGE–C(1.1)	0,007298	0,004544	0,000021	0,015072	0,007135	0,000051
AGE–D(1.3)	0,004444	0,004710	0,000022	0,023031	0,009436	0,000089
AGE–E(1.3)	0,019776	0,016238	0,000264	0,018077	0,008230	0,000068
AGE–B(1.1)	0,002590	0,003742	0,000014	0,014256	0,006855	0,000047
AGE–D(1.1)	0,013010	0,010357	0,000107	0,016054	0,007601	0,000058
AGE–D(1.2)	0,015380	0,007489	0,000056	0,009926	0,008904	0,000079
AGE–E(1.7)	0,005780	0,003815	0,000015	0,018423	0,008000	0,000064
AGE–D(1.5)	0,023198	0,014569	0,000212	0,016905	0,006236	0,000039
AGE–D(1.4)	0,018083	0,018911	0,000358	0,020233	0,017945	0,000322
AGE–B(1.2)	0,019544	0,008907	0,000079	0,030591	0,008091	0,000065
AGE–E(1.4)	0,005059	0,003170	0,000010	0,019441	0,004635	0,000021
AGE–E(1.5)	0,006293	0,004709	0,000022	0,014241	0,007418	0,000055
AGE–A(1.2)	0,018459	0,012320	0,000152	0,026738	0,010217	0,000104
AGE–E(1.6).	0,007098	0,011417	0,000130	0,008318	0,005098	0,000026
AGE–C(1.3)	0,005524	0,003594	0,000013	0,017359	0,006921	0,000048
AGE–B(1.3)	0,006376	0,004784	0,000023	0,032172	0,008989	0,000081
AGE–F(1.2)	0,005663	0,004055	0,000016	0,017649	0,008352	0,000070
AGE–B(1.4)	0,005031	0,005161	0,000027	0,022218	0,007912	0,000063

The correlation matrix shows indicative features which most correlated with each feature. For instance, indicative features are as follows:

AGE–A(1.1) is equally related ($R \approx 0{,}3$) with AGE–E(1.1), AGE–E(1.2), AGE–D(1.3), AGE–B(1.1), AGE–E(1.4), AGE–C (1.3), AGE–B(1.3), AGE–F (1.2), AGE–B (1.4);

AGE–A (1.2) is correlated ($R \approx 0,\ 3$) with the following indicative features: AGE–E (1.2), AGE–E (1.3), AGE–E (1.4), AGE–C (1.3), AGE–B (1.3) and AGE–B (1.4).

AGE–B (1.1) is correlated ($R \geq 0,\ 55$) with the following indicative features: AGE–D (1.3), AGE–E (1.4), AGE–B (1.3) and AGE–B (1.4).

AGE–B (1.2) is equally related ($R \geq 0,\ 5$) з AGE–B (1.1).

AGE–B (1.3) is correlated ($R \geq 0,\ 6$) with the following indicative features: AGE–E (1.1), AGE–E (1.2), AGE–D (1.3), AGE–E (1.7), AGE–E (1.4), AGE–C (1.3), AGE–F (1.2) and AGE–B (1.4).

AGE–B (1.4) is correlated ($R \geq 0,\ 6$) with the following indicative features: AGE–F (1.2), AGE–D (1.3), AGE–E (1.4), AGE–C (1.3), AGE–B (1.1), AGE–E (1.1), AGE–E (1.2).

AGE–C (1.1) is related ($R \approx 0{,}5$) з AGE–B (1.1), AGE–B (1.3), AGE–F (1.2), AGE–B (1.4).

AGE–C (1.2) is equally related ($R \approx 0,\ 45$) з AGE–A (1.1), AGE–B (1.2), AGE–B (1.3).

AGE–C (1.3) is related ($R \geq 0,\ 65$) з AGE–E (1.4), AGE–F (1.2), AGE–B (1.3), AGE–B (1.4).

AGE–D (1.1) is equally related ($R \geq 0,\ 3$) з AGE–F (1.1), AGE–E (1.7) and AGE–D (1.4).

AGE–D (1.2) is equally related ($R \geq 0,\ 2$) з AGE–D (1.5).

AGE–D (1.3) is correlated ($R \geq 0,\ 6$) with the following indicative features: AGE–E (1.1), AGE–E (1.7), AGE–E (1.4), AGE–B (1.3), AGE–B (1.4).

AGE–D (1.4) is equally related ($R \approx 0,\ 3$) з AGE–D (1.1) and AGE–B (1.2).

AGE–D (1.5) is equally related ($R \approx 0,\ 2$) з AGE–D (1.1).

AGE–E (1.1) is related ($R \geq 0,\ 6$) з AGE–D (1.3), AGE–B

(1.3), AGE–F (1.2) and AGE–B (1.4).

AGE–E (1.2) is equally related (R ≥ 0, 55) з AGE–D (1.3), AGE–E (1.7), AGE–E (1.4), AGE–C (1.3), AGE–B (1.3).

AGE–E (1.3) is correlated (R=0, 3) with the following indicative features: AGE–A (1. 2).

AGE–E (1.4) is equally related (R ≈ 0,6) з AGE–E (1.1), AGE–E (1.2), AGE–D (1.3), AGE–B (1.1), AGE–E (1.7), AGE–C (1.3), AGE–B (1.3), AGE–B (1.4), AGE–F (1.2).

AGE–E (1.5) is equally related (R ≥ 0, 5) з AGE–E (1.4), AGE–C (1.3) and AGE–B (1.3).

AGE–E (1.6) is related (R ≥ 0, 2) з AGE–C (1.3).

AGE–E (1.7) is equally related (R ≥ 0, 6) з AGE–D (1.3), AGE–E (1.4), AGE–C (1.3), AGE–F (1.2), AGE–B (1.3).

AGE–F (1.1) is correlated (R=0,378) with the following indicative features: AGE–D (1.1).

AGE–F (1.2) is related (R ≥ 0, 6) з AGE–E (1.1), AGE–E (1.4), AGE–C (1.3), AGE–B (1.4).

As we can see, the indicative features AGE–E (1.4), AGE–C (1.3), AGE–B (1.3) and AGE–D (1.3) have the highest level of correlation.

The next stage of research is the factor analysis. The table of factor loadings shows that 23 indicative features combined to two factors which sufficiently accurately described the results of this research. In this case, the first factor is determined by the following indicative features: AGE–E (1.1), AGE–E (1.2), AGE–D (1.3), AGE–B (1.1), AGE–E (1.7), AGE–E (1.4), AGE–C (1.3), AGE–B (1.3), AGE– F (1.2), AGE–B (1.4). As we can see, the key linguistic-communicative indicators to determine the age of web-user are AGE–B, AGE–E, AGE–F.

These data can be interpreted as the most prominent indicators of Internet communication of adolescents. The second factor is determined by the following indicative features: AGE–D (1.1), AGE–D (1.5) and AGE–D (1.4).

These indicative features form linguistic-communicative indicator AGE–D. This indicator as the key indicator of online-communication of adult is interpreted.

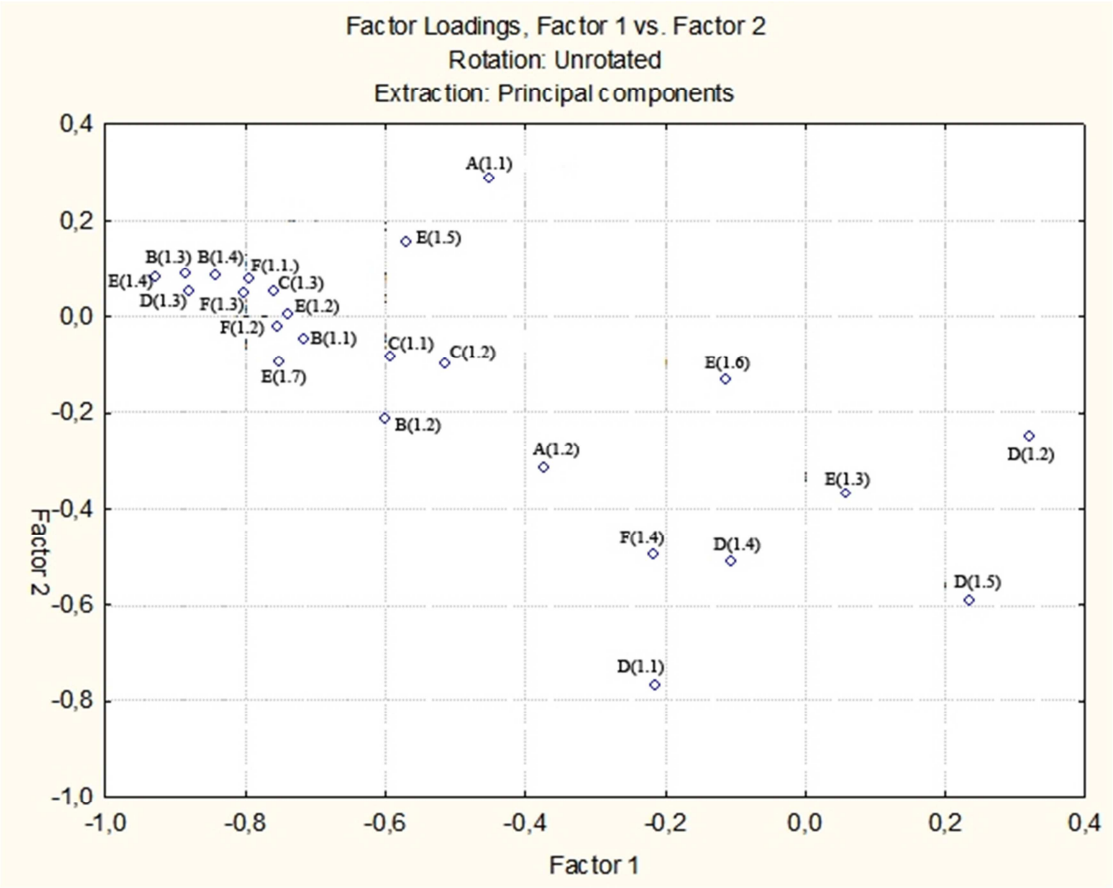

Figure 3. Graph of Factor Loadings: Unrotated rotation

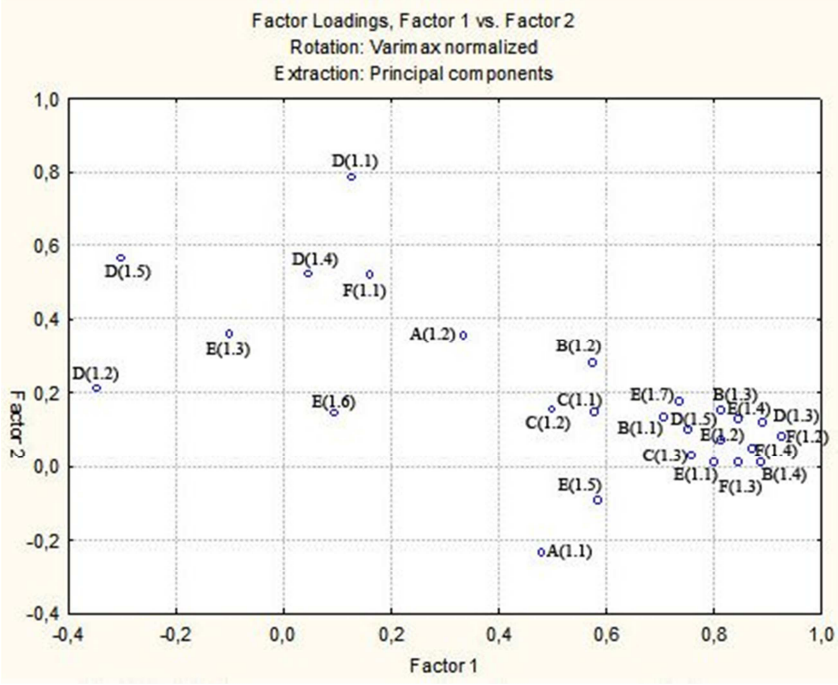

Figure 4. Graph of Factor Loadings: Varimax normalized rotation

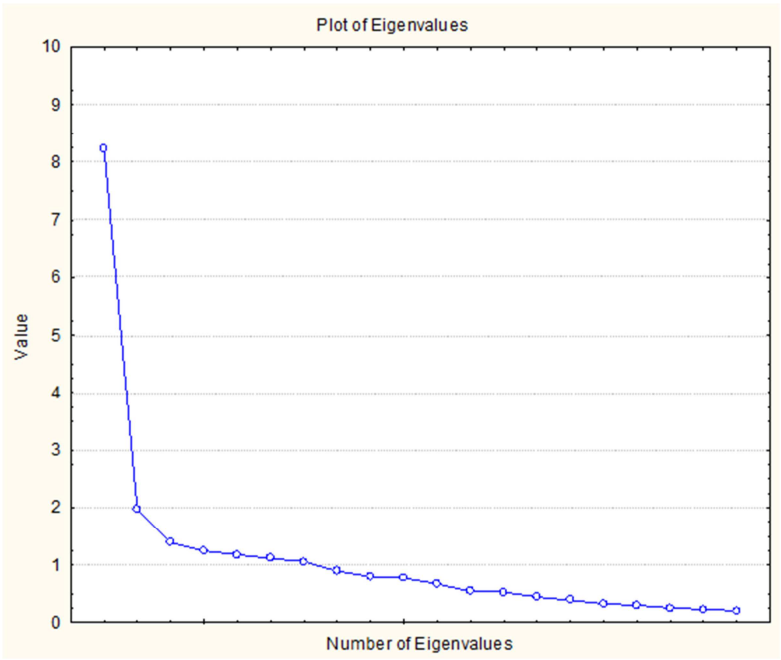

Figure 5. Plot of Eigen value

Thus, the classification of web-forum user into two age groups based on indicative features which are received from the analysis of its content is conducted. Visually, it is shown in the graphs of factor loadings: unrotated rotation (see Figure 3), Varimax normalized rotation (see Figure 4).

The plot of Eigen value (see Figure 5) shows that theoretically for this research is enough select two factors.

The results of classification: After conducting classification of web-forum users in the sample included two monitoring survey (1 adult and 1 teenager) and classified it's again. In this case, the same a priori probability (p=0, 5) of belonging to each cluster is set to both subjects. The both subjects with a posteriori probability p=1 are classified correctly.

4. Conclusions

In this paper the methods of statistical studies the classification of web-forums users: *Lviv. Forum Ridne Misto* and *Rock.Lviv.Ua* are realized. Training sample of web-users

based on the indicative features which by content processing are obtained. The main result of this work is the correct classification of virtual community's users by age. The value of this research lies in verifying basic socio-demographic characteristics of communities' user based on statistical analysis of information track. Every socio-demographic characteristic of virtual community user is determined by analysis of linguistic features in virtual community user's communication. The difference in style of writing posts by virtual community users is the basis for developing effective methods of verifying of personal data in users account. These issues have the greatest influence on efficiency rise of virtual communities functioning and the level of data authenticity in virtual community users' personal profiles. The solution to these problems is possible by using computer-linguistic analysis of web-users' posts. Thus, the importance of statistical methods of virtual community users data verification lied in establishment of mechanisms for collaborative text processing in the global information space and mechanisms of virtual community management (particularly for web-forum administrators and moderators). Received results are basis for development of software [26, 27] for computer-linguistic verification of socio-demographic profile of virtual community user.

References

[1] Fedushko S., et al. (2013) The verification of virtual community user's socio-demographic characteristics profile, Advanced Computing: An International Journal (ACIJ), vol.4, No.3:29-38.

[2] Syerov Yu., et al. (2013) The computer-linguistic analysis of socio-demographic profile of virtual community user, International Journal of Computer Science and Business Informatics (IJCSBI), vol. 4, No 1, 1-13.

[3] Smirnov F. (2006) Art of communication on the Internet. Quick guide. - Williams, 240.

[4] Arinze B., et al. (2002) Some Antecedents and Effects of Trust in Virtual Communities. Journal of Strategic Information Systems, vol.11, 271–95.

[5] Schiano D., et al. (2002) Teen use of messaging media. Proceedings of the ACM Conference on Human Factors in Computing Systems, Minneapolis, NY: ACM Press, 594-595.

[6] Subramanyam K., et al. (2002) The impact of computer use on children's and adolescents' development, Westport, CT: Praeger, 7-30.

[7] Rubio D., Berg-Weger M., Tebb S., Lee E.& Rauch S. (2003) Objectifying content validity: Conducting a content validity study in social work. Social Work Research, 27, 94-104.

[8] Bodnar R. (2004) Sociocultural factors of the teenager sociolect formation. Current studies of foreign languages: Proc. Science. articles, Issue. 2, 142-147.

[9] Bodnar R. (2004) Sociolect of adolescents as an object of linguistics research. Linguistic and conceptual world view: Proc. Science. works. - K. KNU. - № 10. - P.45-50.

[10] Huffaker D. (2004) The educated blogger: Using weblogs to promote literacy in the classroom, 9(6). Available from: www.firstmonday.dk/issues/issue9_6/huffaker/index.html.

[11] Herring S., et al., Women and children last: The discursive construction of weblogs. Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs. Available at: http://blog.lib.umn.edu/blogosphere/women_andchildren.html

[12] Calvert S. (2002) Children in the digital age: Influences of electronic media on development. Westport, CT: Praege, 3-33.

[13] Calvert S., et al. (2002) Identity construction on the internet". Children in the Digital Age: Influences of Electronic Media on Development: Praeger, Westport, CT, 57-70.

[14] Crystal D. (2001) Language and the Internet". Cambridge: Camb. University Press.

[15] Damer B. (2003) Avatars: Exploring and Building Virtual Worlds on the Internet. Longman. Peachpit Press.

[16] Witmer D., et al. (2014) Smile When You Say That: Graphic Accents as Gender Markers in Computer-Mediated Communication, CA: Menlo Park.

[17] Wolf A. (2000) Emotional expression online: Gender differences in emoticon use. CyberPsychology & Behavior, vol.3 (5), 827-833.

[18] Dzyubyshyna-Melnyk N. (2002) Modern slang and modern spoken speech. Scientific notes "KMA", Science, Kyiv, 58-62.

[19] Aksak V. (2006) Communication in the Internet. Certain as two and two, Eksmo.

[20] Virtual Community "Lviv. Forum Ridne Misto". [Internet]. Available from: http://misto.ridne.net.

[21] Web Forum Western rock portal – "Rock.Lviv.Ua". [Internet]. Available from: http://rock.lviv.ua/forum.

[22] Serhyeyenkova O., et al. (2012) Developmental Psychology, K. TSUL, 384.

[23] Hill T., et al. (2007) Statistics methods and applications, StatSoft, Tulsa, OK.

[24] Bureeva N. (2007) Multivariate statistical analysis using PPP "STATISTICA". Training method. material, 112.

[25] Byuyul A., et al. (2002) SPSS: art of information processing. Analysis of statistical data and restore hidden patterns, St. Petersburg. LLC "DiaSoftYuP", 346-367.

[26] Fedushko S. (2014) Development of software for computer-linguistic verification of socio-demographic profile of web-community user. Webology, 11(2), 126. Available at: http://www.webology.org/2014/v11n2/a126.pdf.

[27] Korzh R., Peleschyshyn A., Syerov Y. & Fedushko S. (2014) The cataloging of virtual communities of educational thematic. Webology, 11(1), Article 117. Available at: http://www.webology.org/2014/v11n1/a117.pdf.