Webology, Volume 11, Number 2, December, 2014

Development of a software for computer-linguistic verification of socio-demographic profile of web-community member

Solomia Fedushko
Deputy Dean for Undergraduate Education of Institute of Humanities & Social Sciences, Junior Lecturer of Department of Social Communications and Information Activities, Lviv Polytechnic National University, Ukraine. E-mail: felomia (at) gmail.com

Received October 15, 2014; Accepted December 20, 2014

Abstract

This article considers the current important scientific and applied problem of socio-demographic characteristics validation of virtual community members by computer-linguistic analysis of web-community members' information track. A systematic analysis of the web-members information content and the research of the web-communication specificity of each socio-demographic characteristics value by virtual community content validation for further modeling socio-demographic profiles of web-members are realized. Mathematical models of basic virtual community member socio-demographic characteristics for creating a socio-demographic profile of virtual community member are generated. The method of registration and validation of virtual community member's personal data by checking the maximum amount of virtual community member's data for improving the quality of content and methods of virtual community management is developed. The software for socio-demographic characteristics verification of web-community member, "Socio-demographic profile verifier", is developed, by forming socio-demographic profile of virtual community member that is based on the system building information model of socio-demographic profile of virtual community member for the automation of the verification process of web-custom content.

Keywords

Computer-linguistic analysis; Virtual community; Linguistic and communicative indicator; Socio-demographic marker; Web-community member; Socio-demographic characteristic; World Wide Web

---

Introduction

To take into consideration the trends of virtual communities in a global environment Internet, the research and development of hardware and software community management tools is a priority because the web-community is a common and popular phenomenon, and existing software management tools are imperfect and not integrated. Virtual communities accumulate a huge amount of data. Authenticity of the web-community's users' personal data is an important thing in the successful moderating and managing of web-communities.

The scientific mission of methods and development tools for validation of virtual community members' personal data, including their socio-demographic characteristics, based on computer-linguistic analysis of the content is the current area of the research in computer linguistics. Since without linguistic methods and computer-aided tools this is the hardest task and requires significant time spending for virtual community administrators.

Recent studies have significant practical importance, particularly relevant for qualitative moderation of virtual communities and social and market research.

The main research areas of virtual communities

Analysis of the virtual community performance is the object of scientific researches, including distinction of three main areas (see Figure 1):

- Web usage mining;
- Web structure mining;
- Web content mining.

One of the most important issues of virtual community content analysis today is analysis of web-users' personal data. However, in spite of significant importance for the further development of this research field, methods of analysis are undeveloped in particular personal information on account of virtual communities' users.

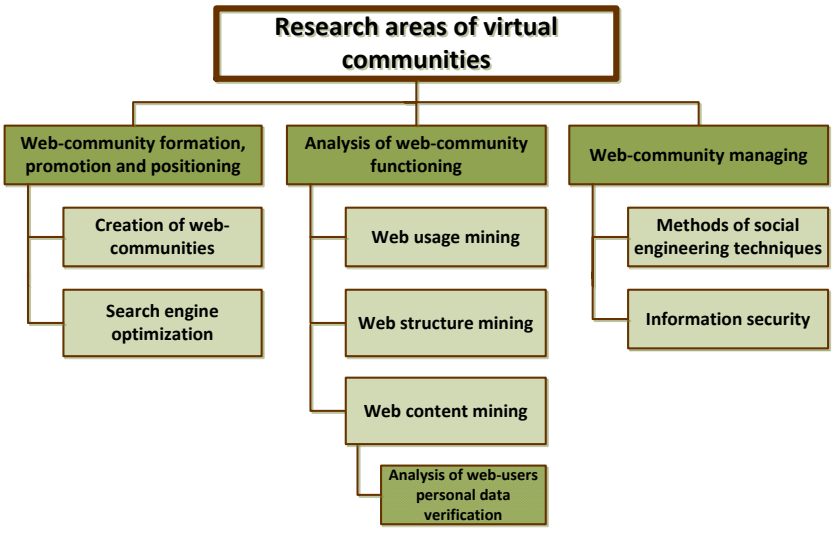

Figure 1. The main research areas of virtual communities

As we see in Figure 1, probability checkup of personal information that is contained in the global system WWW is relevant and important object of the research in the following areas:

- Assessment of the online information reliability (the need of such researches was repeatedly claimed in the works of J. Amsbary and L. Powell, M. Meola, M. Metzger, A. Flanagin, K. Eyal, D. Lemus and R. McCann, T. Johnson and B. Kaye).
- The concept of content reliability (particularly works of A. Anastasi, A. Fitzpatrick, S. Messick, M. Lynn, J. Nunnally, H. Suen, W. Walsh, K. Freeman and J. Spyridakis).
- Content relevance (particularly works of C. Beck and R. Gable).
- The reliability of highly specialized content (works of J. Newhagen and S. Rafaeli).
- The formal review of perception of trust in the information among regular Internet users (in particular works of T. Johnson and B. Kaye).
- The reliability and quality of personal data (S. Park, S. Huh, W. Oh, S. Han, H. Lee, R. Wang, T. Redman and D. Strong).
- Socio-demographic characteristics verification of web-community members of the global environment WWW.

The last one of the researches is currently the least developed. In this scientific direction today are stuck out researches of socio-demographic characteristics verification of users of the global environment WWW, including members of virtual communities. The results of the research in this direction are in demand of a wide range of experts (Korzh, et al. 2014) in the organization and operation of virtual communities (Fedushko, 2010; Peleschyshyn, et al. 2010), as such, that should ensure their performance and efficiency.

This raises an important problem of the new methods and tools development that would have a proper scientific justification, formality, predictable performance and versatility for analyzing the socio-demographic characteristics reliability of the virtual community participants.

For virtual community user data validation (Fedushko & Syerov, 2013), in order to improve virtual communities' management and to improve target techniques in online advertising is enough to analyze such basic socio-demographic characteristics: Age, Sphere of activity, Education level and Gender.

Formation of linguistic and communicative indicator system of virtual community members

Functioning of the formation system of linguistic and communicative indicators involves the content creation and processing of training selection of web-forum members. Scheme of linguistic and communicative indicators formation based on the training selection of web-forum members is shown in Figure 2.

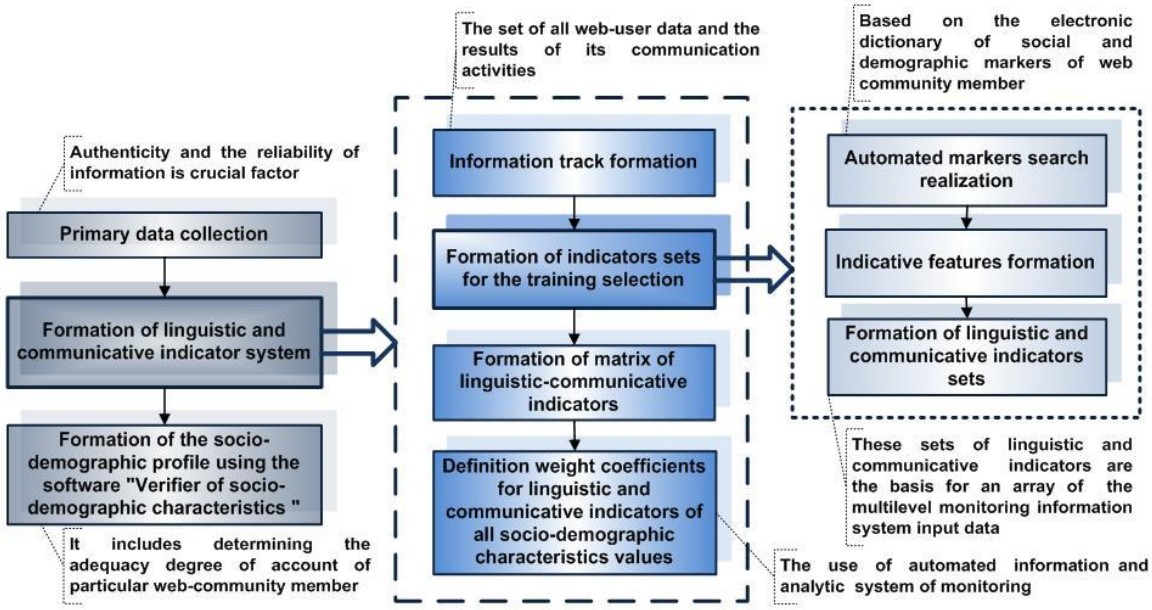

Figure 2. Scheme of formation of linguistic and communicative indicator system based the training selection of web-forum members

Now we will in more detail examine each stage of the algorithm of formation system of linguistic and communicative indicators based the training selection of web-forum members:

Stage I. Primary data collection

Stage II. Formation of linguistic and communicative indicator system

Stage III. Formation of the socio-demographic profile using the software "Verifier of socio-demographic characteristics"

Primary data collection

The primary data collection (Shakhovska, 2011; Shakhovska & Syerov, 2009) is from only credible and reliable virtual community members. Information sources' reliability is crucial for a true and fair result of the research (Fedushko & Bardyn, 2013). For this purpose, is made the collection of primary data from virtual community administrators. To create a training selection only virtual community member with the highest degree of content authenticity is chosen. Administrator or moderator is intimately familiar with virtual community member and Internet communication is tested.

Formation of linguistic and communicative indicator system

At this stage linguistic and communicative indicator sets are formed by automated analysis of information track of virtual community members that is performed in the sequence of steps:

1. **Formation of information track** is carried out according to the information track model of virtual community member.

2. **Formation of indicators sets for the training selection**. According to the structural model of linguistic and communicative socio-demographic characteristic indicators of virtual community

member are chosen virtual community member separation in the following groups according to each investigated socio-demographic characteristics: age, gender, scopes of activity and education. It should be noted that each web-user was carefully chosen for this training selection, considering the virtual community member personal data reliability and reliability of the web-forum member information track.

Also the fact is taken into account that the research results are significantly affected by messages context and discussion topics. In view of this fact, the basis of this study is a diverse sample of user information tracks of all thematic chapters, more than 40 Ukrainian web-forums. Determination of Internet communication features – socio-demographic markers (Peleschyshyn & Fedushko, 2010) is performed by analysis of information track of more than 640 members of Ukrainian virtual communities.

The study equally considered web-forum discussion that arises from a variety of interests and hobbies of young persons and adults, men and women with different levels of education. Computer-linguistic analysis of information track of Ukrainian web-forum members for grammatical, lexical-semantic and lexical-syntactic features laid in analysis more than one specific socio-demographic characteristics value of certain virtual community members. Formation of indicator sets for the training selection is implemented by the following steps:

A. Automated markers search realization

To identify virtual community member socio-demographic characteristics, it is necessary to form linguistic and communicative features (Fedushko, 2011; Fedushko, et al. 2013) a vocabulary of virtual community members Internet communication – socio-demographic markers are determined by the research of phonetic and graphic, formative and lexical-semantic features of virtual space members' speech. Automated markers search of socio-demographic characteristic is done by usage of specialized developed software.

B. Indicative features formation

Indicative features are formed on the basis of common markers of grammatical, lexical-semantic and lexical and syntactic features of virtual community member Internet communication. In order to differentiate virtual community members with socio-demographic characteristics values experts have formed set of gender and age linguistic features, professions and education features of web-users based on:

- researches of scientific theories and ideologies of domestic and foreign leading scientists, linguists, sociologists, psychologists, computer scientists;
- specialized dictionaries (e.g., computer-network jargon dictionary of terms, dictionary of youth slang, etc.);
- content analysis of the Ukrainian virtual communities.

C. Formation of linguistic and communicative indicators sets

The main aim of this process is to consolidate linguistic and communicative indicative features of Internet communication. Formation of linguistic and communicative indicator sets is in grouping indicative attributes in intuitive semantic groups. Visualization of the results is presented in tabular form in the *classification of linguistic and communicative indicators for each value of all socio-demographic characteristics* (Syerov et al. 2013).

Formation of matrix of linguistic-communicative indicators

Based on the linguistic and communicative indicators set experts form the matrix of linguistic and communicative indicators of computer-linguistic analysis of the virtual community content for each value of each socio-demographic characteristics that is defined separately. As a result, for each value of certain socio-demographic characteristics we get a matrix of linguistic and communicative indicators. *The importance of linguistic and communicative is indicated by weight numbers.*

Weight coefficient definition for linguistic and communicative indicator of each socio-demographic characteristic values

Determination of weight coefficients of linguistic and communicative indicators of all socio-demographic characteristic values for each socio-demographic characteristics are done by using information system of multilevel computer monitoring (Golub, 2007). The linguistic and communicative indicator weight numbers of socio-demographic characteristic value are determined by using a multilevel computer information system monitoring. At the stage of the input data array forming of information system multilevel monitoring is processing information tracks of virtual community member in the presence of socio-demographic markers to form linguistic and communicative indicator sets for specific virtual community with the same themes.

The current matrix of linguistic-communicative indicators is an array of input data for information systems' multi-computer monitoring. The input data array of multilevel monitoring information system should meet certain requirements for the synthesis of qualitative multidimensional model and should look like matrices of each marker of linguistic and communicative indicators frequency characteristics in each virtual community member IT, which is the basis for the socio-demographic characteristics models synthesis in information system multilevel computer monitoring.

The most common and popular method for processing such data array is mathematical and statistical method of data processing. However, mathematical and statistical method cannot implement the information system creation for verification of virtual community member personal information.

The example of model of gender linguistic and communicative indicators of virtual community "Lviv. Forum Ridnoho Mista" in the information system of multilevel computer monitoring is shown in Figure 3.

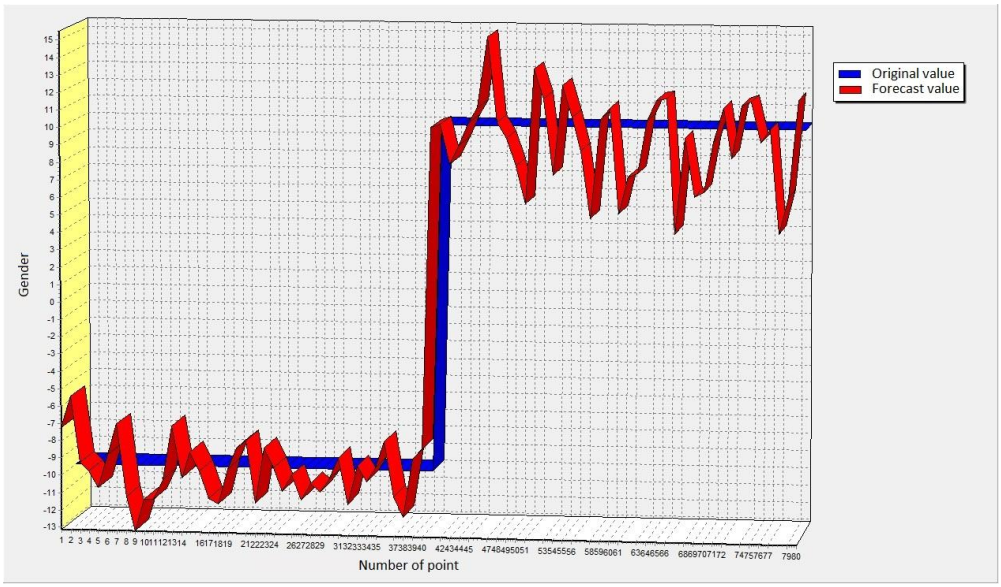

Figure 3. The model of gender linguistic and communicative indicators

This model demonstrates visually affiliation of web-user to a certain socio-demographic characteristic value. The reference rate is formed based on the training set.

It should be noted that the model in the information system of multilevel computer monitoring is synthesized for each virtual community.

Determination of the reliability of the verification process results of socio-demographic characteristics

Reliability of the results of the socio-demographic characteristics verification – is a composite index, which depends on the following parameters: the level of account filling, content topicality, and the relevance of personal data in the account, the technical correctness of filling the account, the administrative authority and virtual community member activity. Reliability of the results of the socio-demographic characteristics verification calculated by the formula:

$$\begin{aligned} \mathrm{RRVer(SDCh)} = k_1 \times \mathrm{Compl}^{UAc} + k_2 \times Actl^{\,UAc} + k_3 \times Actl^{\,Cont} + k_4 \times \\ \mathrm{AdmP}^{User} + k_5 \times TechC^{UAc} + k_6 \times \mathrm{Actv}^{User} + k_7 \times \mathrm{RCB}^{User} + k_8 \times An^{User} \end{aligned} \quad (1)$$

where $k_1, k_2, \ldots, k_8$ – the weight numbers of each parameter of the reliability of the verification process results, which are determined by the member's communicative behavior and virtual community development scenario, with $\sum_i k_i = 1,\ k_i \geq 0$;

$Compl^{UAc}$ – level of account completeness;

$Actl^{\,UAc}$ – relevance of personal data in the account;

$Actl^{\,Cont}$ – level of content topicality;

$AdmP^{User}$ – administrative power;

$TechC^{UAc}$ – level the technical correctness of filling the account;

$Actv^{User}$ – level of user activity;

RCB^{User} – level of compliance with the system of communicative behavior rules of the virtual communities member;

An^{User} – level of anonymity.

As a result, $RRVer(SDCh) \in [0,1]$.

The level of compliance with the system of communicative behavior rules of the virtual community member determined by the formula (2).

$$RCB(User_j) = h^{VCR} \times \left(1 - \sum_i \frac{k_i \times Violation\ (VCR)}{N_j^{VCR}} \right) \qquad (2)$$

where k – weight coefficient of each rule from system of communicative behavior rules of virtual community member, that determined by the development scenario, the web-community mission and the level of harm of breach of virtual community rules $Violation\ (VCR)$ in functioning of the web-community, moreover $\sum_i k_i = 1$, $k_i \geq 0$; h^{VCR} – parameter of rigidity system of communicative behavior rules of virtual community member, which is set by experts (when $h^{VCR} = 1$, than in web-community the system of communicative behavior rules with the highest degree of rigidity is implemented, $h^{VCR} \in [0,1]$); N_j^{VCR} – number of rules that form the system of communicative behavior rules of virtual community member.

Quantity of personal data determines the level of account completeness that is defined as:

$$Compl^{UAc}(User_j) = 1 - \frac{N_{empty}}{N_f} \qquad (3)$$

where N_f – number of filling forms in account,

N_{empty} – number of empty form.

Moreover, $Compl^{UAc}(User_j) \in [0,1]$. Accounts depending on the its value are classified $l_{Compl(User_i)}$.

Activity of virtual community member computes according to the formula 4:

$$Actv(User_i) = k_{Thread} \times \frac{count(Thread(User_i))}{count(Thread)} + k_{Post} \times \frac{count(Post(User_i))}{count(Post)}, \qquad (4)$$

where $count(X)$ – number of elements of the set X;

k_{Thread} *and* k_{Post} – weight coefficient of activity, which determined by expert evaluation considering the development scenario and the web-community mission and the level of harm of communicative behavior rules of virtual community member;

$Thread(User_i)$ – set of all discussions created by the i *-th* member of web-community;

$Thread$ – set of all discussions;

$Post(User_i)$ – set of all messages created by the i *-th* member of web-community;

$Post$ – set of all messages created by the online community members. Moreover, $Actv(User_i) \in [0,1]$.

Table 1. Determination of the reliability level of the result of the socio-demographic characteristics verification

Level of reliability	Value
Reliable result	0,75 < Reliable Result ≤ 1
Ambiguous Result	0, 25 < Ambiguous Result ≤ 0,75
Simulate Result	0 ≤ Simulate Result ≤ 0,25

The reliability of the result of the socio-demographic characteristics verification of virtual community member allows to evaluate the effectiveness of computer-linguistic analysis of virtual community member's content and to construct the socio-demographic profile of virtual community member for the virtual community management and to consider this figure in the virtual community moderating process.

The development of the general software architecture for the socio-demographic characteristics analysis of web-community members

Proposed in the previous chapters' methods is the basis of computer-linguistic analysis software complex of the socio-demographic characteristics reliability of web-community member. In this work the complex architecture reliability check of personal data of virtual community member by computer-linguistic analysis of the socio-demographic characteristic reliability of virtual community member is developed, also the main components of the complex, their functions and technical aspects of implementation is described (see Figure 4).

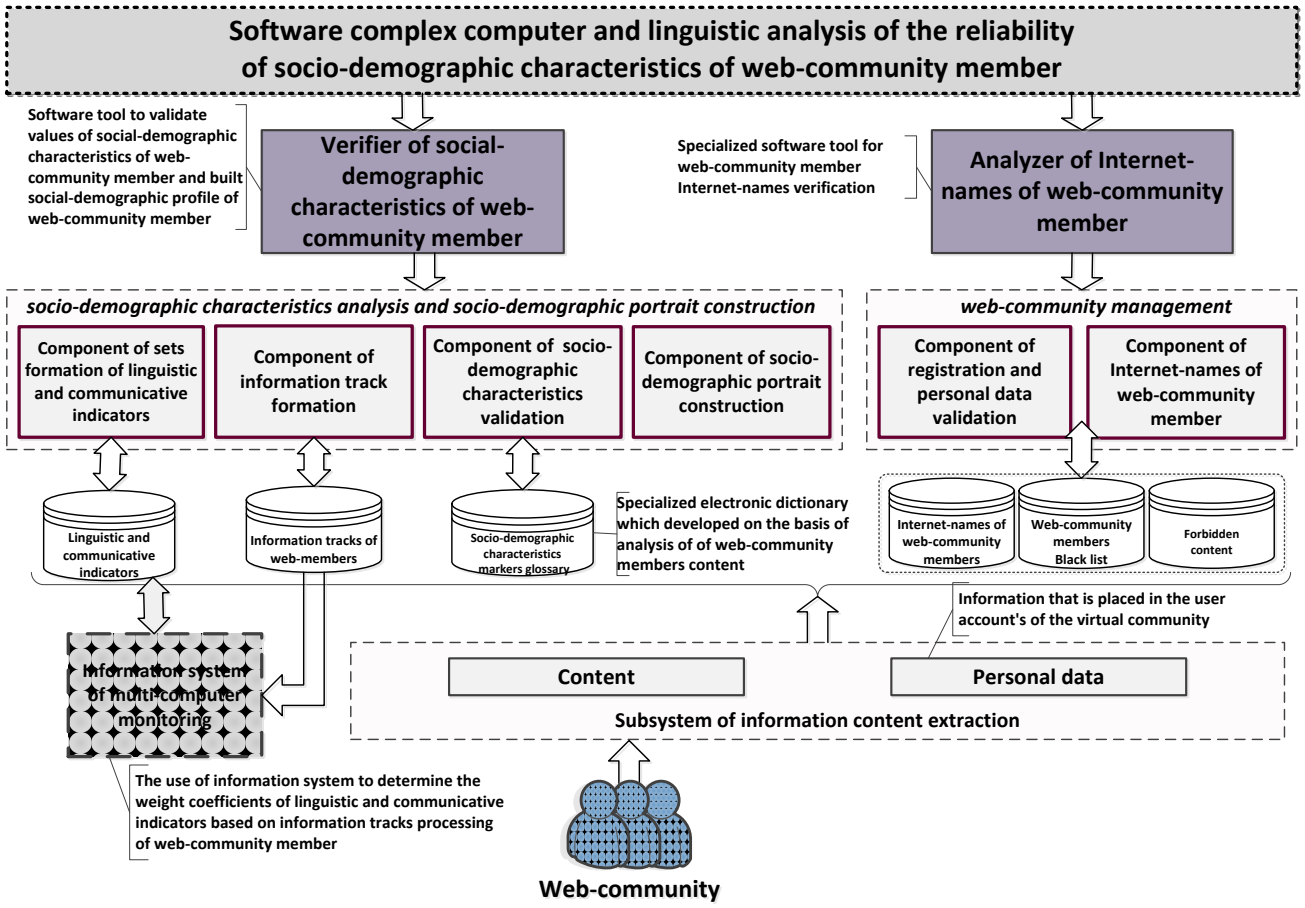

Figure 4. The program complex scheme of computer-linguistic analysis of socio-demographic characteristics virtual community member verification

The development of such software system is divided into the separate stages. The analysis of the subject area and the purpose of the system is the basis for a detailed description of various levels, features and limitations that are imposed on the software system.

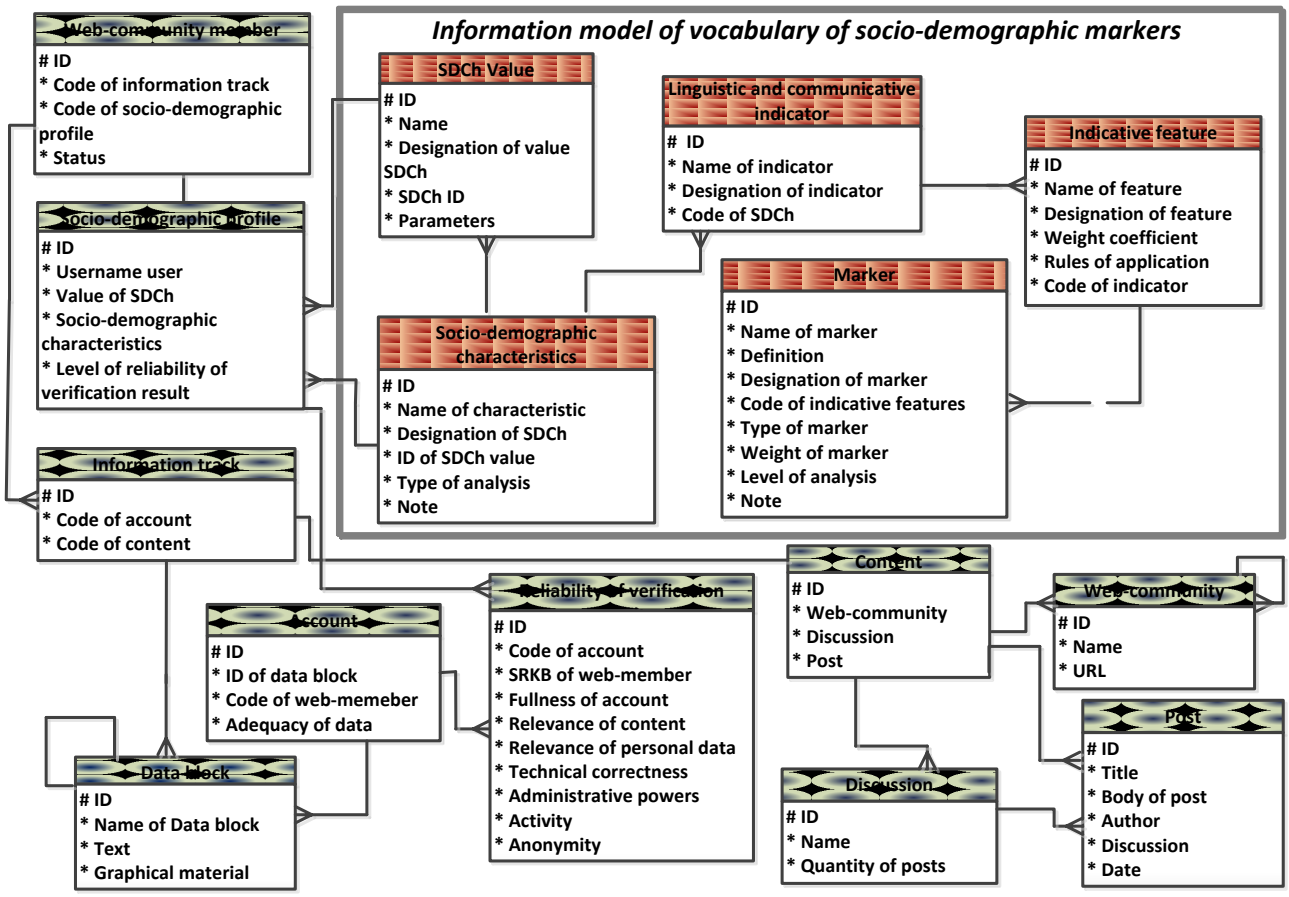

Figure 5. The information model of software for socio-demographic characteristics verification of web-community member "Socio-demographic profile verifier"

The software for socio-demographic characteristics verification of web-community member "Socio-demographic profile verifier" is adapted for one or more languages. This feature is depends on the content vocabulary of socio-demographic markers. The information model of vocabulary of socio-demographic markers in Figure 5 is shown. The level of authenticity of the result of socio-demographic characteristics verification depends on the completeness of filling of vocabulary of socio-demographic markers.

The vocabulary content form linguists, taking into account the content specific of such specialized dictionary for computer-linguistic analysis of socio-demographic characteristics verification of virtual community member. The user interface of software for socio-demographic characteristics verification of web-community member "Socio-demographic profile verifier" in Figure 5 is presented for computer-linguistic analysis of socio-demographic characteristics verification of members of Ukrainian virtual communities.

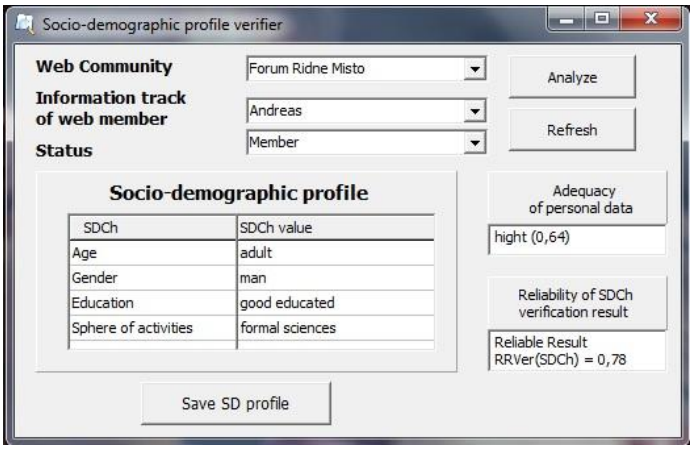

Figure 6. The user interface of software for socio-demographic characteristics verification of web-community member "Socio-demographic profile verifier"

The creation and preservation of socio-demographic profiles of virtual community members is presented on Stage 6 of a scheme of formation of linguistic and communicative indicator system based the training selection of web-forum members (Saving the socio-demographic profile) is.

The results of functioning the software for socio-demographic characteristics verification of web-community member "Socio-demographic profile verifier" is socio-demographic profiles of all web-community members. The socio-demographic profiles of web-forum members "Forum Ridnoho Mista" in Figure 7 are demonstrated.

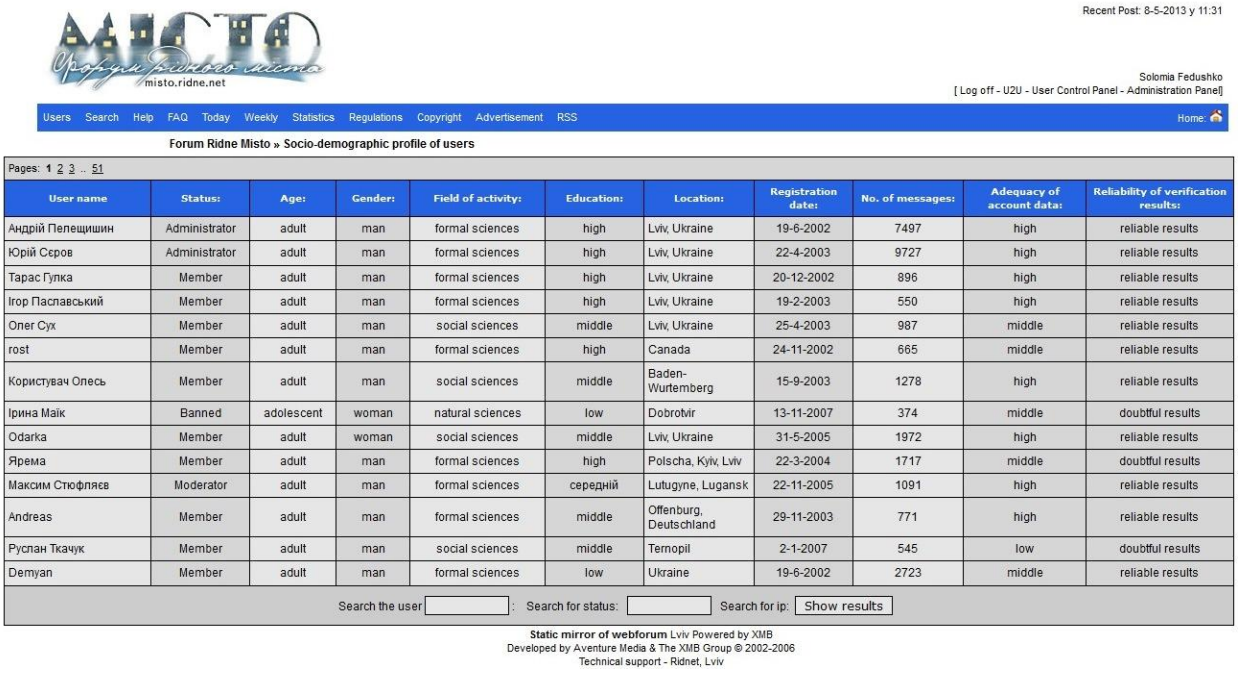

Figure 7. Socio-demographic profile of web-forum members "Forum Ridnoho Mista"

The paper presents a new approach to developing the method of personal data verification of web-users by means of computer-linguistic analysis of web-communities members' information tracks (all information about web-member, which posted on the Internet). Solution of the problem of user data verification is developing and exploiting the software for socio-demographic characteristics verification of web-community member "Socio-demographic profile verifier".

Conclusion

In this paper the important scientific and applied problem has been solved of construction methods and means of basic socio-demographic characteristics validation of virtual community members by computer-linguistic analysis of web-community member's information track. The main scientific and practical results of work are as follows:

- A systematic analysis of the web-members information content and the research of the web-communication specificity of each socio-demographic characteristic value by virtual community content validation for further modeling socio-demographic profiles of virtual community members.
- Mathematical models of basic virtual community member socio-demographic characteristics have been generated for creating a socio-demographic profile of virtual community member.
- The method of registration and validation of virtual community member's personal data by checking the maximum amount of virtual community member data for improving the quality of content and methods of handling virtual community was developed.
- The software algorithmic complex of computer-linguistic analysis of socio-demographic characteristics verification of virtual community member has been created, by forming socio-demographic profile of web-community member that is based on the building information model system of socio-demographic profile of virtual community member for the automation of the verification process of WWW custom content.

References

Fedushko, S. (2010). Disclosure of Web-members Personal Information in Internet. *Proceedings of the Scientific and Practical Conference of Modern Information Technologies in Economics, Management and Education (CITEM-2010)*, Lviv, 2010, 163-165.

Fedushko, S. (2011). Peculiarities of definition and description of the socio-demographic characteristics in social communication. *Journal of the Lviv Polytechnic National University: Computer Science and Information Technology*, Lviv, 694(1), 75-85.

Fedushko, S., & Bardyn, N. (2013). Algorithm of the cyber criminals identification. *Global Journal of Engineering, Design & Technology (GJEDT)*, 2(4), 56-62.

Fedushko, S., & Syerov, Y. (2013). Design of registration and validation algorithm of member's personal data. *International Journal of Informatics and Communication Technology*, 2(2), 93-98.

Fedushko, S., Peleschyshyn, O., Peleschyshyn, A., & Syerov, Y. (2013). The verification of virtual community member's socio-demographic characteristics profile. *Advanced Computing: An International Journal*, 4(3), 29-38.

Golub, S. (2007). The principle of designing multilevel technologies of information modeling. *Journal of Engineering Academy of Ukraine*, 1(1), 28-34.

Korzh, R., Peleschyshyn, A., Syerov, Y., & Fedushko, S. (2014). The cataloging of virtual communities of educational thematic. *Webology*, 11(1), Article 117. Retrieved September 10, 2014, from http://www.webology.org/2014/v11n1/a117.pdf

Peleschyshyn, A., & Fedushko, S. (2010). Gender similarities and differences in online identity and Internet communication. *Proceedings of the International Conference of Computer Science and Information Technologies* (CSIT-2010), Lviv, 195-198.

Peleschyshyn, A., Syerov, Y., & Fedushko, S. (2010). Developing algorithm of registration and validation of personal data on web-community member. *Journal of the Lviv Polytechnic National University: Computer Science and Information Technology*, Lviv, 686(1), 238-244.

Shakhovska, N., & Syerov, Y. (2009). Web-community ontological representation using intelligent dataspace analyzing agent. *X-th International Conference "The Experience of Designing and Application of CAD Systems in Microeletronics" (CADSM-2009)*, Polyana-Svaliava (Zakarpattya), 479–480. Retrieved September 10, 2014, from http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4839885&isnumber=4839735

Shakhovska, N. (2011). Consolidated processing for differential information products. *VII-th International Conference "Perspective Technologies and Methods in MEMS Design"*, Polyana, Ukraine, May 11-14, Lviv, 176.

Syerov, Y., Peleschyshyn, A., & Fedushko, S. (2013). The computer-linguistic analysis of socio-demographic profile of virtual community member. *International Journal of Computer Science and Business Informatics*, 4(1), 1-13. Retrieved September 10, 2014, from http://ijcsbi.org/index.php/ijcsbi/article/view/104/37

Bibliographic information of this paper for citing:

Fedushko, Solomia (2014). "Development of a software for computer-linguistic verification of socio-demographic profile of web-community member." *Webology*, 11(2), Article 126. Available at: http://www.webology.org/2014/v11n2/a126.pdf