International Journal of Computer Science and Business Informatics

IJCSBI.ORG

Determination of the Account Personal Data Adequacy of Web-Community Member

Solomia Fedushko, Yuriy Syerov, Andriy Peleschyshyn and Korzh Roman

Social Communications and Information Activities Department,
L'viv Polytechnic National University,
Ukraine, L'viv, S. Bandera Street 12

ABSTRACT

This article considers the current problem of investigation and development of data verification of virtual community member by means of computer-linguistic analysis of web-members socio-demographic profiles for web-community member identification. The algorithm of formation system of lingvo-communicative indicators based the training selection of web-forum members is designed. This algorithm includes the formation of matrix of linguistic-communicative indicators and lingvo-communicative indicator weight coefficient determination. The computer-linguistic analysis of web-community members' information tracks is realized and verification of lingvo-communicative indicators of gender, age and sphere of activities of web-community member is established. Based on these results of the investigation the software algorithm of web-community members' socio-demographic profiles verification is designed and software "Verifier of the socio-demographic profile" is exploited. The account personal data adequacy of web-community member is determined by means of computation of the measure of the adequacy of personal account data. The calculation method of reliability of the result of the socio-demographic characteristics verification of web-community member is developed for constructing the socio-demographic profile of the web-member for web-community management.

Keywords

Socio-demographic profile, Software, Marker, Web-community member, Personal data, Validation.

1. INTRODUCTION

The most important issues of web-communities content analysis today is analysis of web-users' personal data [1, 2]. 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 web-communities' users. Development of web-community management software is a priority issue because the web-community is a popular and mass service at WWW, and existing software management tools are imperfect and not integrated. Authenticity of web-community user personal data is an important thing in the successful managing of web-communities.

The scientific task of validation method development for web-communities members' personal data, including their socio-demographic characteristics, based on computer-linguistic analysis of the informational 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 web-communities administrators.

2. BACKGROUND STUDY

The verification of personal data that is contained in the global system WWW is relevant and important research object in the following areas:

- Assessment of the online information reliability.
- The concept of content reliability, content relevance and reliability of highly specialized content.
- The formal review of perception of trust in the information among regular Internet.
- The personal data reliability and quality.
- Verification of socio-demographic users' characteristics in WWW.

The results of the research in the last scientific direction – of socio-demographic characteristics verification of users of the global environment WWW [3, 4] – are in demand of a wide range of experts in the organization and operation of web-communities, 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 web-communities members.

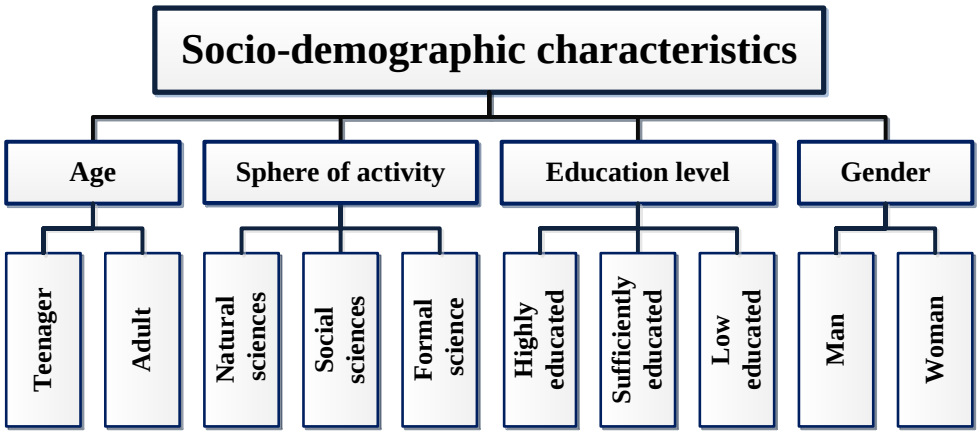

Figure 1. The structure studied by social-demographic characteristics of a web-user

For web-community user data validation, in order to improve web-communities management [5] and to improve target techniques in online advertising [6] is enough to analyze such basic socio-demographic characteristics (see. Figure 1).

2.1 The reasons of socio-demographic characteristics validation:

Obtaining data from real users of social networking pages ("Facebook", "Twitter", "Habrahabr", "Google+", etc.), the popular online newspapers and magazines, dating sites ("LoveUA", "FindLove.in.UA", “12 Kisses”, “Dating foreigners website”) and web-forums (Ukrainian forum of programmers “Replace.Org.Ua”, “Domivka.Net: Ukrainian Forum”, “The first site of the Ukrainian community in Italy”) is urgent task for web-community administrators, police, private detectives (services that have been increasingly needed by users of global environment Internet) and individual user of any information resource.

Verification and setting of web-users’ personal data who performs illegal actions that are offensive to opponents from psychological, financial, legal angle, in particular, who levies blackmail, sends hate mails, spreads false and aspersion information and conduct other electronic bullying.

2.1.1 Age

Age category is chosen for specified age validation in web-community member due to a number of important factors: the presence of real online threats to Internet users with the age of 6 to 17 years (these include disclosure of personal confidential information, access to content that do not meet age peculiarities and adversely affect the physical and mental health of the child, online abuse, internet marketing crimes, etc.) and the need for screening age group of children who have applied or already are web-community member, that is intended only for adult web-space users.

The characteristic features of a web-identity are self-expression and experimentation with the aim to make a definite impression on the users of the web-community and this is peculiar especially to growing age.

The problem of age differentiation is that the penetration rate of children is increasing in web-communities, which are intended for adult users' communication, and this destroys communicative atmosphere of community and vice versa, children are increasingly claiming to be adult members. This situation can lead community administrator to criminal charges [7]. An example of web-communities for teenagers online communication is: "Student Forum|UNIVER-SITY", "chat for teens", chat "teenager", "chat for teens", "Just chat" and for adults "Our anecdote with Pepper" and others.

2.1.2 Education

Many community administrators require web-community members to follow certain conventions of web-communication. Web-communication convention depends on the objective goal and projected scenarios of the web-community owners. An important factor in the possibility of participating in a web-community is a high literacy level of web-community member [8], level of higher education and skills.

The method of screening illiteracy and with low education level web-community member will significantly reduce the time and financial costs, and managers' efforts to moderate web-community. In order to screen illiterate web- member it required to classify all members by the level of literacy, which will help to determine the likely level of each web-community member education. In the task automation research of detecting errors in the text are only considered character errors. Error analysis and data consolidation that were conducted by scientists, allows us to offer the errors typology and to classify members in the order of education level.

2.1.3 Scope of activity

A web-member in a greater or lesser degree can belong to several areas of interest, as the impact of the content creation in web-communities serve a direction of education, professional activity and range of interests in spare time. However, the result of the analysis is to determine the scope of activity of web-community member. In the general scope of activities are classified as science fields in the following areas: Natural sciences; social sciences; formal science. This division is based on analysis of web-community members Internet communication with different scopes of activities.

In web-community moderation that is intended for communication between members of certain areas, there many questions appear about level of professionalism, ethical principles, so filtering of web-users by scope of activities is a necessary and important task of community administration (government and policy portals) to improve the web-community position.

2.1.4 Gender

In order to avoid gender conflicts in web-communities that are intended for women (L'viv Women's Forum “Cult of Beauty”, "Girls chat", Forum "L'viv-mama" et al.) and for men ("Man Forum", "Anti-feminine site" etc.) administrators of these communities need to enter a strict gender division.

"Web-gender change", that is a web-user creation identity of the opposite sex, it is common on the Internet and is caused by the following factors as the impact of cultural gender stereotypes, expression of homosexual tendencies or transsexual or diffuse gender identity. So-called web-sex change is more peculiar to men due to several reasons desire to control and manipulate others, for women it is easier to get help and attract attention,

desire to power over other men, the study of the relationship between sexes and get new experience of Internet communication.

2.1.5 The geographic location

Locating web-user with no verification data on computer IP-address of the web-community member using special software tools (CNGeoip, GeoLite Country, GeoLite City, IpGeoBase, GeoIp etc.) that can determine the belonging of IP-address to the level of countries and cities is the aim of geographic location identification of web-member.

3. LINGVO-COMMUNICATIVE INDICATORS FORMATION

The formation system of lingvo-communicative indicators involves the content creation and processing of training selection of web-forum members.

The algorithm of formation system of lingvo-communicative indicators based the training selection of web-forum members consist of these stages:

I. Primary data collection

II. Lingvo-communicative indicators formation

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

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 - is performed by analysis of information track of more than 640 members of Ukrainian web-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 are more specific to one particular socio-demographic characteristic value of certain web-community members.

The experts set of gender and age linguistic features, professions and education features of web-users are formed based on:

- Researches of scientific theories and ideologies of domestic and foreign leading scientists, linguists, sociologists, psychologists, computer scientists;

- Specialized dictionaries;
- Content analysis of the Ukrainian web-communities.

The main aim of this process is to consolidate lingvo-communicative indicative features of Internet communication. Formation of lingvo-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 lingvo-communicative indicators for each value of all socio-demographic characteristics.

3.1 Formation of matrix of linguistic-communicative indicators

Based on the lingvo-communicative indicators set experts form the matrix of lingvo-communicative indicators by computer-linguistic analysis of the web-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 lingvo-communicative indicators:

$$LKI^{(SDCh,Vc)} = \begin{pmatrix} Ind_{1,1}^{(SDCh,Vc)} & \cdots & Ind_{1,i}^{(SDCh,Vc)} & \cdots & Ind_{1,N\_Vl(SDCh,Vc)}^{(SDCh,Vc)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ Ind_{j,1}^{(SDCh,Vc)} & \cdots & Ind_{j,i}^{(SDCh,Vc)} & \cdots & Ind_{j,N\_Vl(SDCh,Vc)}^{(SDCh,Vc)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ Ind_{N\_Ind(SDCh,Vc),1}^{(SDCh,Vc)} & \cdots & Ind_{N\_Ind(SDCh,Vc),i}^{(SDCh,Vc)} & \cdots & Ind_{N\_Ind(SDCh),N\_Vl(SDCh,Vc)}^{(SDCh,Vc)} \end{pmatrix}$$

where *N_Vl* - a feature that for each socio-demographic characteristics identifies a number of socio-demographic characteristic values; *N_Ind* - a feature that for each socio-demographic characteristics identifies a number of lingvo-communicative indicators of this socio-demographic characteristics value.

Each matrix row is a vector of lingvo-communicative indicators of some socio-demographic characteristics:

$$Ind^{(SDCh,Vc)} = \left(Ind_{1,1}^{(SDCh,Vc)} \quad \cdots \quad Ind_{N\_Ind(SDCh,Vc),i}^{(SDCh,Vc)} \quad \cdots \quad Ind_{N\_Ind(SDCh),N\_Vl(SDCh,Vc)}^{(SDCh,Vc)} \right)$$

The vector of certain value socio-demographic characteristic *SDCh* of specific certain web-community *Vc*:

$$LKI^{(SDCh,Vc)} = \begin{pmatrix} Ind_{1,1}^{(SDCh,Vc)}(U) \\ Ind_{j,1}^{(SDCh,Vc)}(U) \\ Ind_{N\_Ind(SDCh,Vc),1}^{(SDCh,Vc)}(U) \end{pmatrix}$$

Table 1 Tabular representation of functions: *N_Vl & N_Ind*

SDCh	*N_Vl*	*N_Ind*
Age	2	6
Gend	2	12
Edu	3	7
Sphere	3	11

To calculate the distance from the reference socio-demographic characteristics value to each possible socio-demographic characteristic value of atomic each web-community member we take as a basis the formula for determining the Euclidean distance:

$$\rho_j^{(k)}(Value, User) = \sqrt{\sum_{i=1}^{N\_Ind(SDCh,k)} \left(Ind_{i,j}^{(SDCh,Vc)} - Ind_{i,j}^{(SDCh,U)}\right)^2 * w_i^{(SDCh)}}$$

where $k \in 1 \ldots N\_Vl(SDCh,Vc)$; $w_i^{(SDCh)}$ - weight coefficient of particular lingvo-communicative indicator of particular value of socio-demographic characteristic.

As a result we take such value of socio-demographic characteristic which corresponds $\rho^* = min(\rho_k)$ to the maximum value $Value(U)$. Moreover, the matrix $LKI = (Ind_{ij})$ is universal for all values of socio-demographic characteristics of a particular web-community, for which are synthesized models. Depending on the subject and the type of web-community, the model is synthesized for each socio-demographic characteristic values using automated information system monitoring.

The weight coefficients of lingvo-communicative indicators are presented in the vector:

$$W^{(Vl,SDCh)} = \left(w_1^{(Vl,SDCh)} \quad \ldots \quad w_j^{(Vl,SDCh)} \quad \ldots \quad w_{N\_Ind(SDCh,Vc)}^{(Vl,SDCh)}\right)$$

The weight coefficient vector of indicators *SDCh* socio-demographic characteristics - *Vl* value, obtained as a result of automated information system monitoring.

The importance of lingvo-communicative is indicated by weight coefficients.

3.2 Lingvo-communicative indicator weight coefficient determination

The weight coefficient determination for lingvo-communicative indicators of all socio-demographic characteristic values for each socio-demographic characteristic is completed by using information system of multilevel computer monitoring. At the stage of the input data array forming of information system multilevel monitoring is processing information tracks

of web-community member for the presence of socio-demographic markers to form lingvo-communicative indicator sets for specific web-community with the same themes.

The matrix of lingvo-communicative indicators is an array of input data for information system's multi-computer monitoring.

The input data array of multilevel monitoring information system should meet certain requirements for the synthesis of qualitative multidimensional model and must be matrices of each marker of lingvo-communicative indicators frequency characteristics in each web-community member information track. It is the basis for the socio-demographic characteristics models synthesis in information system multilevel computer monitoring.

4. SOFTWARE FOR SOCIO-DEMOGRAPHIC PROFILE VERIFICATION OF WEB-COMMUNITY MEMBER

The complex architecture reliability check of web-community member personal data by computer-linguistic analysis of the socio-demographic characteristics reliability of web-community member is developed, also is described the main components of the complex, their functions and technical aspects of implementation.

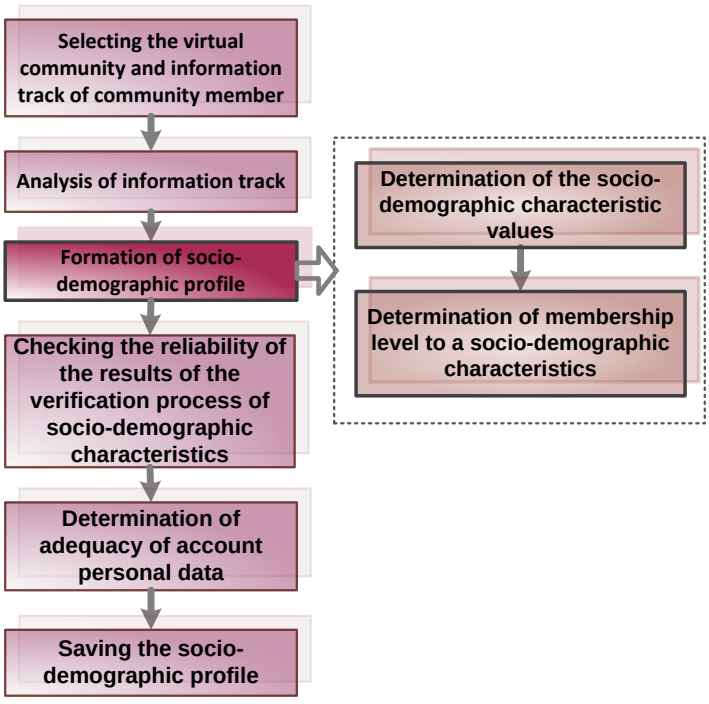

Figure 2. The scheme of functioning of "Socio-demographic profile verifier"

The developed system of sets of web-member Internet communication lingvo-communicative indicators is the basis for software of test socio-demographic characteristic values of web-communities members - "Verifier of socio-demographic profile". The scheme of software algorithm "Verifier of socio-demographic profile" is described in Figure 2. The algorithm of "Verifier of socio-demographic profile" is presented in the following stages:

Step 1. Selecting the web-community and information track of member.

Step 2. Analysis of information track.

Step 3. Formation of socio-demographic profile. This step included two stage: determination of the socio-demographic characteristic value and determination of membership level to a socio-demographic characteristics.

Step 4. Checking the reliability of the results of the verification process of socio-demographic characteristics.

Step 5. Determination of adequacy of personal account data.

Step 6. Saving the socio-demographic profile. The result of this step is the creation and preservation of socio-demographic profiles of web-members.

5. DETERMINATION OF THE ACCOUNT PERSONAL DATA ADEQUACY OF WEB-COMMUNITY MEMBER

Adequacy of account personal data - is the characteristic of account personal data which indicates results reliability degree of the verification process of socio-demographic characteristics of a particular web-community personal data that is specified in correspondent account, that is, the determination of the account personal data veracity.

The measure of the adequacy of account personal data - a certain level of probability that is analyzed by computer-linguistic analysis of web-community member account to reference web-community member account based on real and relevant information about web-community member.

The difference between 1 and $\rho_j^{(k)}(Value, User)$- is the distance between the reference socio-demographic characteristics value and *k*-th web-community member atomic socio-demographic characteristic value that is determined as an adequate account personal data of *k*-th user.

$$\mu_j^{(k)}(Value, User) = 1 - \rho_j^{(k)}(Value, User)$$

where $\rho_j^{(k)}(Value, User)$ - the distance from the reference socio-demographic characteristics value to each possible socio-demographic characteristics value of atomic *k*-th web-community user:

$$\mu_j^{(k)}(Value, User) = 1 - \sqrt{\sum_{i=1}^{N\_Ind(SDCh,k)} \left(Ind_{i,j}^{(SDCh,Vc)} - Ind_{i,j}^{(SDCh,U)}\right)^2 * w_i^{(SDCh)}}$$

where $k \in 1 \ldots N\_Vl(SDCh,Vc)$. Moreover, $\mu_j^{(k)}(Value, User) \in [0,1]$.

This vectoring method consists of the data transformation in the vector form that will allow determining the extent of similarity between the socio-demographic characteristic values. The similarity measure value between the socio-demographic characteristics value and control vector (the value that is determined in training selection of web-forums members for each set of socio-demographic characteristic values) indicates a web-community member identity to a certain socio-demographic characteristic value. The analysis results vary according to the web-community specificity. To greater ratio value corresponds the more important linguistic-communicative indicator for the verification of socio-demographic characteristics in particular web-community.

6. RELIABILITY OF THE VERIFICATION PROCESS RESULTS

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

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 web-community member activity. Reliability of the results of the socio-demographic characteristics verification calculated by the formula:

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

where $k_1, k_2, \ldots, k_8$ – the weight coefficients of each parameter of the reliability of the verification process results, which are determined by the member's communicative behavior and web-community development scenario, with $\sum_i k_i = 1$, $k_i \geq 0$; $Compl^{UAc}$ – the level of account filling; $Actl^{UAc}$ – the relevance of personal data in the account; $Actl^{Cont}$ – the level

of content topicality; $AdmP^{User}$ – the administrative authority; $TechC^{UAc}$ – the level the technical correctness of filling the account; $Actv^{User}$ – the level of activity; RCB^{User} – the level of compliance with the web-community member rules; An^{User} – the level of anonymity. As a result, $RRVer(SDCh) \in [0,1]$. Determination of the reliability level of the result of the socio-demographic characteristics verification: 0,75 < Reliable Result ≤ 1; 0, 25 < Ambiguous Result ≤ 0,75; 0 ≤ Simulate Result ≤ 0,25.

7. CONCLUSIONS

Determination of the account personal data adequacy of web-community member is the important scientific and applied problem. The construction methods and means of basic socio-demographic profiles validation of web-communities members by computer-linguistic analysis of web-members information track is solved this problem.

The system of lingvo-communicative indicators involves the content creation and processing of training selection of web-forum members is formed. Weight coefficients of lingvo-communicative indicators are determined. The matrix of linguistic-communicative indicators is formed. The software algorithm of web-community members' socio-demographic profile verification – "Verifier of socio-demographic profile" is designed.

Computer-linguistic analysis of information track of Ukrainian web-forum members for grammatical, lexical-semantic and lexical-syntactic features are more specific to one particular socio-demographic characteristic value of certain web-community members. The research results are significantly affected by messages context and discussion topics. 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 socio-demographic markers is performed by analysis of information track of more than 640 members of Ukrainian web-communities. The study equally considered web-forum discussion that arises from a variety of interests of young persons and adults, men and women with different levels of education.

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

8. REFERENCES

[1] Fedushko, S., Syerov, Yu., 2013. Design of registration and validation algorithm of member's personal data, *International Journal of Informatics and Communication Technology*, Vol.2, No.2, pp. 93-98.

[2] Shakhovska, N., 2011. Methods of customer data processing using intelligent agent of data sources structure determination. *Actual Problems of Economics*, Vol. 7(120), pp. 338-346.

[3] Fedushko, S., Peleschyshyn, O., Peleschyshyn, A., Syerov, Yu., 2013. The verification of web-community member's socio-demographic characteristics profile, *Advanced Computing: An International Journal*, Vol.4, No.3, pp. 29-38.

[4] Syerov, Yu., Peleschyshyn, A., and Fedushko, S., 2013. The computer-linguistic analysis of socio-demographic profile of web-community member. *International Journal of Computer Science and Business Informatics*, Vol. 4, No. 1, pp. 1-13.

[5] Shakhovska, N. and Syerov, Yu., 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"*, Polyana-Svaliava, Ukraine, pp. 479–480.

[6] Shakhovska, N., 2011. Consolidated processing for differential information products. *VII-th International Conference "Perspective Technologies and Methods in MEMS Design"*, Polyana-Lviv, Ukraine, p. 176.

[7] Fedushko, S., Bardyn, N., 2013. Algorithm of the cyber criminals identification. *Global Journal of Engineering, Design & Technology*, Vol. 2, No. 4, pp. 56-62.

[8] Korzh, R., Peleschyshyn, A., Syerov, Yu., and Fedushko, S., 2014. The cataloging of virtual communities of educational thematic. *Webology*, 11(1), pp. 1-16.

This paper may be cited as:

Fedushko, S., Syerov, Y., Peleschyshyn, A. and Roman, K., 2015. Determination of the Account Personal Data Adequacy of Web-Community Member. *International Journal of Computer Science and Business Informatics, Vol. 15, No. 1, pp. 1-12*.