Determination of Web Data Adequacy

Syerov Stanley
The M&R Companies
Chicago, USA
syerov@yahoo.com

Fedushko Solomiia
Кафедра СКІД
НУ "Львівська політехніка"
felomia@gmail.com

This article considers the current problem of investigation of data verification of virtual community member. 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.

Keywords: account, personal data, adequacy, web-community, member.

INTRODUCTION

Obtaining data from real users of web pages, online newspapers and magazines, dating sites and web-forums is urgent task for web-community administrators, police, private detectives and individual user of any information resource. Verification of web-users' personal data, who performs illegal actions, is actual task. The main aim of this process is to consolidate lingvo-communicative indicative (LCI) features of Internet communication. Formation of LCI sets is in grouping indicative attributes in intuitive semantic groups. Visualization of the results is presented in tabular form in the classification of LCI for each value of all socio-demographic characteristics (SDCh).

WEB DATA ADEQUACY DETERMINATION

Based on the LCI set experts form the matrix of LCI by computer-linguistic analysis of web-community content for each value of each SDCh that is defined separately. As a result, for each value of certain SDCh we get a matrix of LCI:

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

where *N_Vl* - a feature that for each SDCh identifies a number of SDCh values; *N_Ind* - a feature that for each SDCh identifies a number of LCI. Each matrix row is a vector of LCI:

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

The vector of certain value *DCh* of specific certain web-community *Vc*:

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

The formula for determining the Euclidean distance is used to calculate the distance from the reference SDCh value to each possible SDCh value of atomic each web-community member:

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

where $k \in 1 \ldots N\_Vl(SDCh, Vc)$; $w_i^{(SDCh)}$ - weight coefficient of particular LCI. As a result we take value of SDCh, which corresponds $\rho^* = min(\rho_k)$ to maximum value $Value\ (U)$. Matrix $LKI = (Ind_{ij})$ is universal for all SDCh of particular web-community. The weight coefficients of LCI are presented in vector:

$$W^{(Vl,SDCh)} = \begin{pmatrix} w_1^{(Vl,SDCh)} \ldots w_j^{(Vl,SDCh)} \ldots \\ \ldots w_{N\_Ind(SDCh,Vc)}^{(Vl,SDCh)} \end{pmatrix}$$

The weight coefficient vector of indicators *SDCh* - *Vl* value, obtained as result of automated information system monitoring. Weight coefficient determination for LCI of all SDCh values for each SDCh is completed.

Matrix is array of input data for information system multi-computer monitoring.

Adequacy of account personal data - is the characteristic of account personal data which indicates results reliability degree of the verification process of SDCh 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-member account to reference web-member account based on real and relevant information. The difference between 1 and $\rho_j^{(k)}(Value, User)$- is the distance between the reference SDCh value and *k*-th web-member atomic SDCh value that is determined as adequate data of *k*-th user.

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

where $\rho_j^{(k)}(Value, User)$ is the distance from the reference SDCh value to SDCh:

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

Where $k \in 1 \ldots N\_Vl(SDCh, Vc)$. Moreover, $\mu_j^{(k)}(Value, User) \in [0,1]$. This vectoring method consists of data transformation in vector form that will allow determining extent of similarity between SDCh values.

Analysis results vary according to web-community specificity.

CONCLUSION

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 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 web-forums.

REFERENCES

[1] S. Fedushko, “Development of a software for computer-linguistic verification of socio-demographic profile of web-community member”, Webology, vol. 11 (2), , 2014. www.webology.org/2014/v11n2/a126.pdf

[2] S. Fedushko, Yu. Syerov, and R. Korzh, “Validation of the user accounts personal data of online academic community,” IEEE XIIIth Intern. Conf. “Modern Problems of Radio Engineering, Telecommunications and Computer Science”, 2016, pp. 863-866.

[3] S. Fedushko, "Development of verification system of socio-demographic data of virtual community member", Radio Electronics Computer Science Control, 3, p. 87-92, 2016.

[4] Y. Syerov, S. Fedushko, Determination of Development Scenarios of the Educational Web Forum, XI International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT), p.73-76, 2016.