CHAID USER SEGMENTATION ANALYSIS OF BEHAVIOURAL BIOMETRIC FEATURES FOR ONLINE PLATFORMS

Daniel PETCU[1], Dan Alexandru STOICHESCU[2]

The emergence and the large scale use of mobile devices and the increase in use and ease of access to internet technologies create the opportunity and the necessity for the owner of online platforms to analyse efficiently the internet behavioural aspect of users. This interaction between man and terminal described also as HCI (human computer interaction) can be analysed in order to obtain quality information and to further adapt non-intrusive techniques. Behavioural biometrics techniques are implemented and used on large scale by different vendors at present.

This paper proposes an implementation for CHAID Analysis of behavioural biometric features as an instrument for online platforms in order to achieve better categorization of active users in order to come forward with better demanded solutions. The implemented extraction algorithm obtains the behavioural biometric features which are used in the CHAID analysis. The obtained output results and the CHAID algorithm are computed using related models IBM SPSS Statistics 20. The results are further discussed.

Keywords: Behavioural biometrics, decision trees, Chi-square automatic interaction detection (CHAID), information security, traits, online platform, PHP

1. Introduction

Identity authentication and verification [1] is in our society a major factor in daily routine starting from logging on computer machine, unlocking and using a mobile smart-phone, gaining access into an office and using one's bank account to perform transactions [2].

Biometric technologies provide authentication or identification of individuals based on physiological or behavioural features. It is therefore necessary for these traits to be easily identifiable and verifiable. The use of such features depends essentially on premises that reside in permanence character and distinctiveness metrics. In terms of describing such technologies there can be

[1] PhD. student, Faculty of Electronics, Telecommunications and Information Technology University POLITEHNICA of Bucharest, Romania, e-mail:danpetcu33@yahoo.com

[2] Prof., Faculty of Electronics, Telecommunications and Information Technology University POLITEHNICA of Bucharest, Romania, e-mail: stoich@elia.pub.ro

stated that physiological biometric features use what we are as individuals and behavioural biometric features use what we do. Analysing the user-interaction perspective, physiological biometrics can be affected by skill of the user in the acquisition part of the traits and it implies certain cooperation level [3]. Furthermore there are cases with specific health issues and physical disabilities that are to be noted. Based on these aspects, one can state that in terms of acceptance, behavioural biometrics technologies are accepted with ease by the user due to the fact that it does not introduce delays in operation and it can be implemented silently with no other user requirements in terms of cooperation, in most of the online platforms, along with desired system [4].

Looking in terms of numbers of research papers, in field of biometrics it can be observed and estimated, that techniques which use the Internet user interaction in online environment are fewer compared with the traditional physiological approaches. In these traditional approaches, most of the biometric systems implementations use extracted traits with specific predominance from fingerprint and iris [5]. Comparing with traditional physiological approaches in field of biometrics one can observe that in current interconnected informational society also known as networked society, an increase in access and mobility to internet connected devices or computers along with the lack of dedicated hardware for acquisition of biometric data, demanding lower costs of implementation, also poses a challenge for segmentation and extracting relevance behavioural traits [2].

Though these behavioural techniques based on human-device interaction offer insufficient data and robustness for accurate authentication, verification of users and segmentation can be achieved based on biometric profiles of users [6]. Looking from the level of human-device interaction, these techniques can be classified as direct and indirect techniques. The first group of direct techniques relies on the basic interaction between man and input device such as keyboard, mouse, haptic-device or any other device that is controlled by the user with mechanic interaction. The second group of indirect techniques makes use of the human-device interaction on strategy level, utilization of interface, utilization of lexical corpus, and time metrics. This group of indirect biometric techniques offers a variety and wide perspective for monitoring key behavioural events useful in the construction of behavioural user profiles and time tree interface interaction [6].

Due to nature of use in every online platform, the user has to interact in specific way to access specific functions or to find certain information and this requires the user to go through a route in the interface.

This paper presents such an approach from this second group of indirect biometric techniques and makes use and establishes an analysis of behavioural traits extracted from an online platform. In the following section this paper

presents the CHAID methodology. Finally the paper presents the related work implementation and the delivered results obtained with IBM SPSS Statistics 20.

2. CHAID Methodology

The methodology of classification based on trees of classification is used in prediction of appurtenance for the statistical units in classes of a categorical dependent variable, based on the measurements performed on one or multiple predictors. Such classification trees are an important technique utilized in segmentation and data mining [7].

Chi-square automatic interaction detection, also known as CHAID is a method of segmentation and prediction based on tree classification variance analysis. It evaluates all the values of potential predictor features using a statistical significance test criterion and it is commonly used when a dependent categorical variable exists and plenty of independent categorical variables are used in the analysis [8].

In every point of the CHAID analysis, it is identified the best predictor from a subgroup of unities. The main aspect is to merge the nearest categories from the multitude of independent variables with many categories [7].

The delicate part is to merge those categories in order to obtain increased homogeneities. This process is achieved in two steps. The first step implies the search for the categories from de same independent variable that will merge. In order to achieve the mergence of categories, a sequence of stages is done for all the independent variables.

1. It is constructed a bi-dimensional table of frequencies for the dependent variable with the independent variable of whose categories we expect to fuse.
2. It is calculated the χ^2 used for independence statistics for each pair of categories and the associated p value defined as the probability that the independence hypothesis is accepted.

$$\chi 2 = \sum_{j=1}^{s}\sum_{i=1}^{r} \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}} \approx \chi 2_{(r-1)(s-1)} \quad (1)$$

where n_{ij} are the frequencies and $n_{i\bullet}$, $n_{\bullet j}$ are the totals on lines and columns.

3. It is determined the pair with the highest p value. In case that this value is higher than a predefined significance level (i.e. 0.005), the 2 categories will fuse.

4. In case there are only 2 categories that can fuse, the p value is adjusted with the use of *Bonferroni multiplier[11][7]* :

$$\chi 2 = \sum_{i=0}^{r-1} (-1)^i \frac{(r-i)^c}{i!(r-i)!} \tag{2}$$

where c is the number of independent variables and r is the number of categories that were merged. If this value is greater than the predefined level of significance, the two categories will fusion.

These stages in this first fusion step are performed for all independent variables. Finally, in the second step, the variables used for the ramification of the constructed tree, are determined. Here the node is branched based on the variable for which the lowest p value was obtained, adjusted after fusions, lesser than the significance predefined level.

The process stops when an ending criterion is reached. Such an ending criterion can be defined when no independent variable has a significant p value or when a subgroup contains few observations.

The output of this analysis is a tree from where decision and related information can be interpreted.

3. Web-platform interface implementation

The structural bloc diagram, followed in the implementation of the web platform is found in Fig. 1:

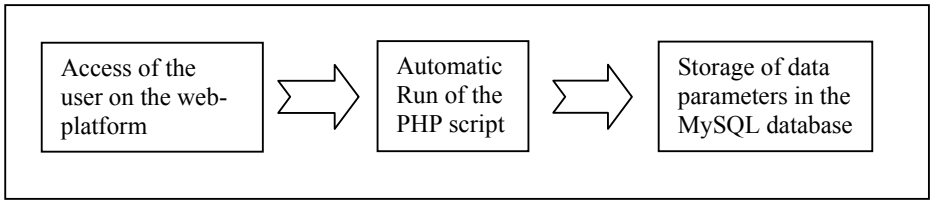

Fig.1. Web-platform interface bloc diagram

For each user that accesses the web-platform, the implemented script in the system will verify if there is a new user or if is a user that has already visited the platform based on the existence of the specified cookie in the PHP script. If the user is new to the system then in the browser used for accessing the web-platform, a new cookie of permanent persistence with the “User Key” unique identifier will be created. Each time the user accesses the web-platform the system automatically allocates a specific session key “Session Key”. Therefore a user that

accesses the web-platform will have the same "User Key" stored in the Cookies but the "Session Key" for each user is generated automatically.

This automatic session generation is performed when the user accesses a web page from the platform and it will have session life-time duration of 30 minutes.

The developed script is written in PHP (PHP: Hypertext Pre-processor), a programming language used on large scale in the development of online web applications. The information related to the user access is stored in MySQL database. This implementation is chosen taking into account the accessibility and large scale convenience in order for the end user ease of access to the web-platform.

Each web page of the platform is provided with a line that calls the developed implemented tracking function needed to extract the biometric traits. The input parameter of this function is the accessed web page. All the other needed parameters are calculated based on previous accesses of the web-platform or automatically taken. Time spent on each page is computed by calculating the difference between the access time on current page and the access time on the previous web page. For monitoring the number of pressed backspaces and the time of insertion of desired keyword in the search functionality of the web platform we have developed a JavaScript to enable access for related parameters. The developed JavaScript code is run in real-time by the browser and it has reduced size in order not to affect the performances of the web-platform. Also the web-platform extracts the IP address of each user along with all specific biometric and non-biometric parameters. The functionality diagram of the system integrated in the web platform is specified in the Fig 2.

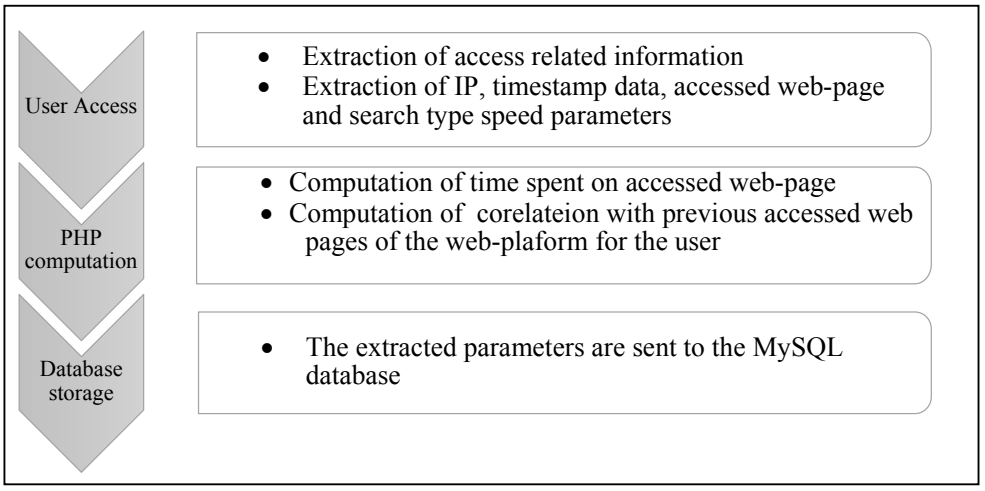

Fig. 2. Web-platform functionality diagram

4. Analysis of Biometric traits and related work

The biometric traits are extracted and post-processed in order to obtain maximum relevant information to take a decision in terms of behavioural point of view.

In practical terms the behavioural traits [9] that are used as measured parameters in the analysis are coded in Table 1.

Table 1

Traits and Behavioural Traits used

Trait Name	Definition	Unit of Measure	Type
User Key	Unique generated user key allocated for each user that accesses the web platform	N/A	Non Behavioural
Session Key	Defined as Unique generated session key by the web platform	N/A	Non Behavioural
IP	IP address of user	Standard IP v4 notation	Non Behavioural
Total Number of unique session keys per user	Defined as the number of distinct unique session keys per user	Decimal number	Behavioural
Total Number of pages per all sessions per user	Defined as the number of pages per all sessions per user	Decimal number	Behavioural
Δ web-space	Defined web-space distance between source pace and next browsed page of the platform	a-dimensional	Behavioural
Δ time	Defined as difference between registered time of the incoming source page inside of platform and current accessed platform web page	Seconds	Behavioural
Search function used	Defined as the usage of search functionality by each user that access the web platform	Yes/No	Behavioural
Search function type speed	Defined as the time that user writes desired key-word in search functionality of the web platform	Seconds	Behavioural
Nr of usages of Backspace in Search function	Defined as the number of Backspace key usages for each session for unique user that accesses web platform	Decimal number	Behavioural

The desired framework of this analysis is to conclude the relevance of the proposed behavioural traits in keen relation to the non-behavioural traits and

evaluate them in the proposed CHAID methodology in order to obtain the best decision tree with minimum standard deviation error.

The "Δ web-space" behavioural trait is constructed using Levenshtein distance [10] between the current source web-page name and the incoming web-page name measured per individual user key and individual session.

Both pages are named accordingly with desired keyword relevance in significance to the accessed web pages. Thus this methodology for the Levenshtein distance in information theory and computer science is a string metric for measuring the difference between two sequences. Usually this distance between two keywords is best known as the minimum number of single-character alterations.

For these alterations literature refers to insertions, deletions or substitutions required to change one word into the other [10].

Before the CHAID conducted analysis an important aspect that reflects the unique patterns for each user behaviour that access the web platform can be found in the diagram in figure Fig 3.

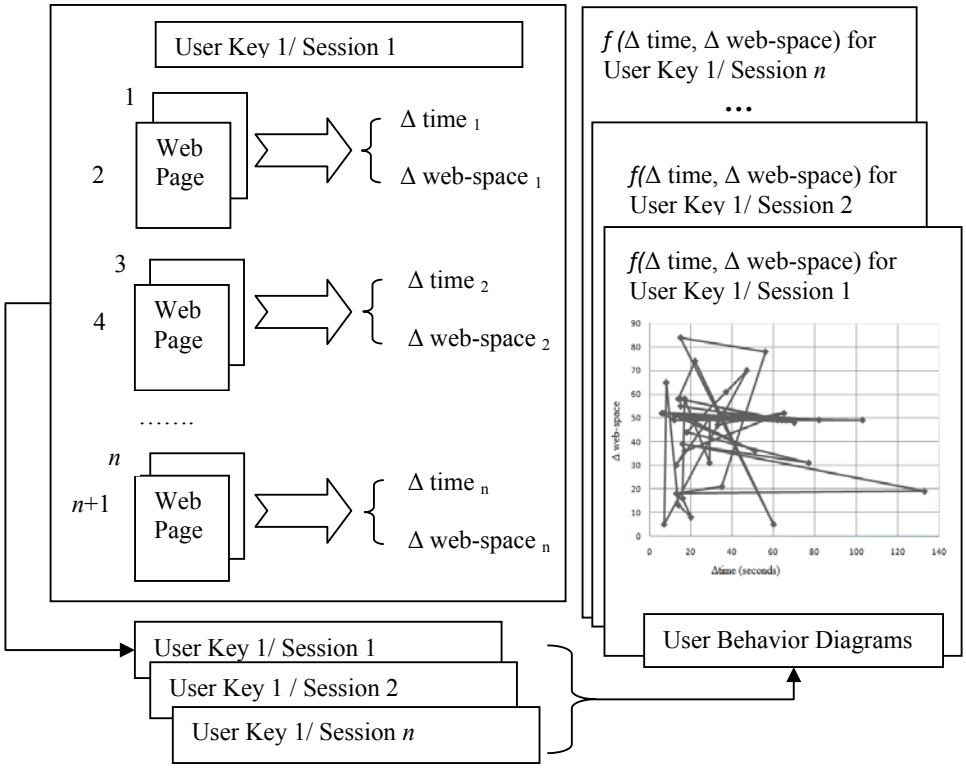

Fig.3. User behaviour Diagrams in Δ(time) per Δ(web-space) domain

The above user behaviour diagrams per sessions can be normalized on both scales and correlation factor parameters between different users can be computed. In this analysis for better results we can include a filtering step where same user accessing platform from different IP's in less than 30 seconds will be excluded and treated as machine entities.

The implemented CHAID methodology [11] uses the following input traits as in Table 2.

Table 2

Behavioural Traits used in CHAID analysis

Trait Name	SPSS Label	Segmentation Details	SPSS Measure	Role
Total Number of unique session keys per user	Unique_S	Very low – 1 session Low – 2 sessions Medium – 3 to 7 sessions High – 8 to 12 sessions Very High – more than 12	Nominal	Input
Total Number of accessed web pages per all sessions per user	Total_pages_acc	1 – 1 web pages [2-10] – between 2 and 10 [11-20] – between 11 and 20 [21-50] – between 21 and 50 >50 – more than 50	Nominal	Input
Δ web-space	Delta_web_sp	0 [1-30] [31-60] [61-90] >90	Nominal	Input
Δ time	Delta_time	<15s [15s – 5min] (5min – 15 min] (15min – 30min] >30min	Nominal	Input

The implementation of the CHAID analysis is performed in IBM SPSS software [11]. In the CHAID growing method the chosen dependent variable is "Unique_S" with following independent variables "Total_pages_acc", "Delta_web_sp", "Delta_time". The maximum tree depth is set to 3 with minimum cases in parent node set to 2000 and minimum cases in child node set to 1000.All three independent variables specified are included in the model with no rejected variable.

5. Experimental Results

The data volume is pre-processed according to described CHAID analysis methodology and uses up to 27000 records extracted during 1 month of web-platform running.

The resulting CHAID diagram is a graphic representation of the tree model. This tree diagram is shown in Fig 4.

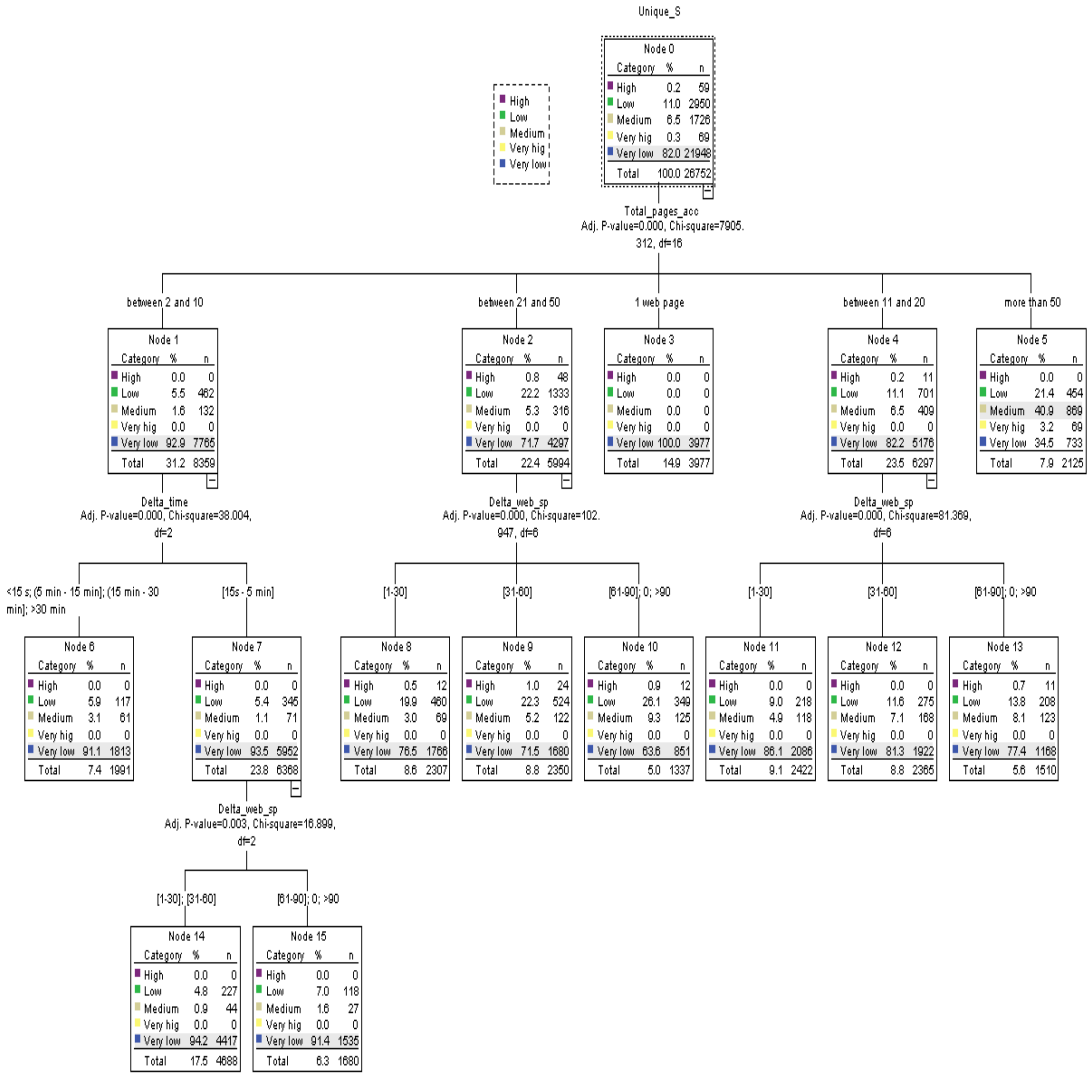

Fig.4. CHAID Tree diagram for behavioural traits model

Using the CHAID method, the level for the total number of accessed web pages on all sessions per user is the best predictor of the rating for the total number of unique session keys per user.

The defined „more than 50" category of the total number of accessed web pages on all sessions per user is the only significant predictor of total number of unique session keys per user. From all users accessing the web-platform in this category, 40.9% have „medium" level for the total number of accessed web pages on all sessions per user. Thus on the resulted diagram there are no child nodes below the above mentioned node, this is considered a terminal node.

The node related to „1 web pages" category from the total number of accessed web pages on all sessions per user is also observed as a terminal node in the diagram. In this category all the users accessing the web-platform have „very low" total number of accessed web pages on all sessions per user.

For the „between 11 and 20" and „between 21 and 50" categories from the total number of accessed web pages on all sessions per user, the next best predictor is Δ web-space behavioural trait metrics.

For the „between 2 and 10" category from the total number of accessed web pages on all sessions per user, the next best predictor is Δ time behavioural trait metrics.

For the „between 2 and 10" category from the total number of accessed web pages on all sessions per user users accessing web-platform categorized in [15 s – 5 min] Δ time, the CHAID model includes one more predictor, Δ web-space. Over 94% of those users segmented in „[1-30]" or „[31-60]", Δ web-space classes have a „very low" rating of total number of unique session keys per user, while 91% of these users segmented in „0", „[61-90]" and „>90" have „very low" unique session rating.

6. Conclusions

From all the proposed behavioural biometric traits that were extracted by the developed web-platform, in this paper we have selected to include in the CHAID analysis the following: the „total number of unique session keys per user", the „total number of accessed web pages per all sessions per user", the „Δ web-space" and „Δ time".

After computing and generating the experimental results in the analysis there are obtained the decision model parameters, chi-square value of the CHAID method, degrees of freedom (df) and significance level (Sig.) for the split. For most practical purposes of the analysis, the interest resides in the significance level, which is less than 0.0001 for all splits in this model.

A measure of the tree's predictive accuracy of this current analysis is the risk estimate and its standard associated error.

For categorical dependent variables, the risk estimate is the proportion of cases incorrectly classified after adjustment for prior probabilities and misclassification costs.

For the used categorical dependent variables, the computed output classification table shows the number of correctly and incorrectly classified cases for each category of the total number of unique session keys per user, dependent variable.

The risk and classification tables show in terms of evaluation, how well the computed model works.

The obtained risk estimate of 0.174 indicates that the category predicted by the behavioural traits model is wrong for 17.4% of the cases. So the "risk" of misclassifying a user accessing the platform is approximately 18%. The results in the computed classification table are consistent with the risk estimate and it shows that the model classifies approximately 82.6% of the users correctly.

The results obtained conclude that constructed parameters and the related behavioural traits present significance for future related works and we intend to extend analysis to all parameters related in table 1.

Acknowledgement

The work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/134398.

R E F E R E N C E S

[1]. *Anil K. Jain, Arun A. Ross and Karthik Nandakumar*, "Introduction to Biometrics", Springer, 2011, ISBN 978-0-387-77325-4

[2]. *Giles Hogben*, "Behavioural Biometrics Final Report", European Network and information Securty Agency (ENISA), Feb 2010

[3]. *Arun A. Ross, Anil K. Jain, Karthik Nandakumar*, "Handbook of multibiometrics", Springer, 2006, ISBN: 978-0-387-22296-7

[4]. *R Yampolskiy, V. Govindaraju*, "Behavioral Biometrics: a Survey and Classification". International Journal of Biometrics (IJBM), 2008, **vol.** 1, no. 1, pp. 81-113.

[5]. *Andy Adler, Richard Youmaran and Sergey Loyka*, "Pattern Analysis and Applications: Towards a measure of biometric feature information", Springer, **vol** 12, no. 3, Sept 2009, pp 261-270

[6]. *Kenneth Revett*, "A Bioinformatics Based Approach to Behavioural Biometrics", Conference on Frontiers in the Convergence of Bioscience and Information Technologies, 2007, IEEE, ISBN 978-0-7695-2999-8

[7]. *Press Laurence I., Rogers Miles S.and Shure, Gerald H.*, "An interactive technique for the analysis of multivariate data", Behavioral Science, 1969, **vol.** 14, pp. 364–370

[8]. *Ahmed Hamza Osman and Naomie Salim*, "An Improved Semantic Plagiarism Detection Scheme Based on Chi-squared Automatic Interaction Detection", 2013 International

Conference on Electrical and Electronics Engineering (ICCEEE), pp 640-647, Aug. 2013, ISBN 978-1-4673-6231-3

[9]. *Hamid Banirostam, Elham Shamsinezhad and Touraj Banirostam*, "Functional Control of Users by Biometric Behavior Features in Cloud Computing", 4th International Conference onIntelligent Systems Modelling & Simulation (ISMS), Jan 2013, ISBN 978-1-4673-5653-4

[10]. *Rane, S and Wei Sun*, "Privacy preserving string comparisons based on Levenshtein distance", IEEE International Workshop on Information Forensics and Security (WIFS), 2010, US, ISBN: 978-1-4244-9078-3

[11]. *Snježana Pivac*, "Detection and solving of regression modeling problems in SPSS", MIPRO Proceedings of the 33rd International Convention 2010, Croatia, May 2010, pp 914 – 919, ISBN: 978-1-4244-7763-0