Journal of Theoretical and Applied Information Technology
15th February 2012. Vol. 36 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

APPLICATION OF K-NEAREST NEIGHBOUR PREDICTOR FOR CLASSIFYING TRUST OF B2C CUSTOMERS

[1]MEHRBAKHSH NILASHI, [2]KARAMOLLAH BAGHERIFARD, [3]OTHMAN IBRAHIM, [4]NASIM JANAHMADI, [5] BAHMAN PANJALIZADEH, [6]MOUSA BARISAMI

[1] Dept. of Computer Engineering, Islamic Azad University, Roudsar and Amlash Branch, Roudsar, Iran
[2] Dept. of Computer Engineering, Islamic Azad University, Yasooj branch, Yasooj, Iran
[3] Assoc. Prof., Faculty of Computer Science and Information Systems, UTM, Skudai, Malaysia-81310
[4] Dept. of Computer Engineering, Islamic Azad University, Roudsar and Amlash Branch, Roudsar, Iran
[5] Islamic Azad University, Roudsar and Amlash Branch, Roudsar, Iran
[6] Dept. of Computer Engineering, Islamic Azad University, Roudsar and Amlash Branch, Roudsar, Iran

E-mail: nilashidotnet@yahoo.com, karam_bagheri@yahoo.com, othmanibrahim@utm.my, ,janahmadi.nasim@hotmail.com, b.panjalizadeh@gmail.com,barisamy.lahijan@gmail.com

ABSTRACT

K-nearest neighbor (k-NN) classification is one of the most fundamental classification methods and should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data. In addition, nearest neighbor analysis is a method for classifying cases based on their similarity to other cases. In this paper using k-NN method some factors that affect on customer trust in online transactions, were classified. Raw data gathered from customers when they were buying as customer in B2C websites. One questionnaire was developed and data was gathered from online customers. After organizing data, k-NN method was applied and desired results were obtained. Results showed that in which positions customer can trust to B2C websites and which factors are more significant. Accordingly, in this paper k-NN enable us to predict role of factors on trust level in five levels.

Keywords: *K-NN, Trust, B2C, Security, Customer.*

1. INTRODUCTION

The K-nearest neighbor (k-NN) method for classification is one of the most straightforward approaches to classifying objects which are represented as points defined in some feature space. Despite the simplicity of KNN the performance it achieves on a number of pattern recognition tasks indicates that it remains competitive as a classification method [1] [10] [11] [12] [13] [15].

The k-nearest neighbor (k-NN) method is a common hot deck method, in which k donors are selected from the neighbors (i.e., the complete cases) such that they minimize some similarity measure [2]. In the k-NN method, missing values in a case are imputed using values calculated from the k nearest neighbors, hence the name. The nearest, most similar, neighbors are found by minimizing distance function, usually the Euclidean distance, defined as [3]:

Distance between two scenarios can be computed using some distance function d(x,y) , where x,y are scenarios composed of N features, such that $x=\{x_1,x_2,....x_N\}, y=\{y_1,y_2,.....y_N\}$.

Two distance functions are discussed in this summary:

$$d_A(x,y)=\sum_{i=1}^{N}|x_i-y_i| \qquad (1)$$

Euclidean distance measuring:

$$d_E(x,y)=\sum_{i=1}^{N}|x_i^{\,2}-y_i^{\,2}| \qquad (2)$$

Because the distance between two scenarios is dependant of the intervals, it is recommended that resulting distances be scaled such that the arithmetic mean across the dataset is 0 and the standard deviation 1. This can be accomplished by replacing the scalars x, y with x', y' according to the following unction:

$$x'=\frac{x-\bar{x}}{\sigma(x)} \qquad (3)$$

Where x is the unscaled value, $\bar{x}$ is the arithmetic mean of feature x across the data set (see

Journal of Theoretical and Applied Information Technology
15th February 2012. Vol. 36 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Equation 4), $\sigma(x)$ is its standard deviation (see Equation 5), and x' is the resulting scaled value. The arithmetic mean is defined as:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (4)$$

Also the standard deviation can be computed as follows:

$$\sigma(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \qquad (5)$$

A measure in which to determine the distance between two scenarios, we can simply pass through the data set, one scenario at a time, can be established and compare it to the query scenario.

Data set as a matrix D = N*P, containing P scenarios s_1, s_2, ..., s_P can be represented, where each scenario s_i contains N features. A vector O with length P of output values O = {O_1, O_2, ..., O_P} accompanies this matrix, listing the output value Oi for each scenario Si .

KNN can be run in these steps:

1- Store the output values of the M nearest neighbors to query scenario q in vector {r^1 ...,r^m} by repeating the following loop Al times:

A) Go to the next scenarios' in the data set, where i is the current iteration within the domain 1,…, P}

B) If q is not set or q < d (q, s^i): q ←d(q, s^t),t←o^i.

C) Loop until we reach the end of the data set (i.e. i = P)

D) Store q into vector c and t into vector r

2- Calculate the arithmetic mean output across r as:

$$\bar{r} = \frac{1}{M}\sum_{i=1}^{M} r_i \qquad (6)$$

3- Return $\bar{r}$ as the output value for the query scenario q [14].

The use of Euclidean distance as similarity measure is recommended by Strike et al. [4] and Troyanskaya et al. [5]. The k-NN method does not suffer from the problem with reduced variance to the same extent as mean imputation, because when mean imputation imputes the same value (the mean) for all cases,k-NN imputes different values depending on the case being imputed.

In machine learning, it was developed as a way to recognize patterns of data without requiring an exact match to any stored patterns, or cases. Similar cases are near each other and dissimilar cases are distant from each other. Thus, the distance between two cases is a measure of their dissimilarity.

Cases that are near each other are said to be "neighbors." When a new case (holdout) is presented, its distance from each of the cases in the model is computed. The classifications of the most similar cases – the nearest neighbors – are tallied and the new case is placed into the category that contains the greatest number of nearest neighbors. We can specify the number of nearest neighbors to examine; this value is called k. The pictures show how a new case would be classified using two different values of k. When k = 5, the new case is placed in category 1 because a majority of the nearest neighbors belong to category 1. However, when k = 9, the new case is placed in category 0 because a majority of the nearest neighbors belong to category 0.Figure 1 shows the effect of changing k on classification.

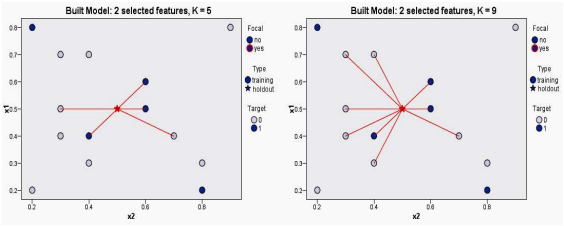

Figure 1. The Effects of Changing K on Classification

Nearest neighbor analysis can also be used to compute values for a continuous target. In this situation, the average or median target value of the nearest neighbors is used to obtain the predicted value for the new case.

2. TRUST IN E-COMMERCE

Web sites provide people with a convenient way to disseminate information. Although the Internet has expanded tremendously during the past decade, the high reluctance for using eCommerce and on-line business activity still remains. Based on the previous surveys, two major problems are web site security and the lack of Web site trust. Although most eCommerce Web sites provide some forms of secured payment method, it doesn't guarantee that the Web sites will gain better credibility. Based on many previous research and studies, there are many other significant factors which could influence the trust level of the Web sites layout. W3 Trust Model

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(W3TM) proposed trust value calculation from meta data provided on the Web sites [6], however, it does not include the analysis on other important factors such as contents and layout analysis and more importantly the ability to automatically recommend trust-assessment improvement. At present, there are no effective tools for evaluating trust assessment on Web sites. An effective trust assessment tool must be able to identify and determine the trustworthiness level of the Web sites correctly. In addition, it should provide recommended information based on trust assessment in order for Web masters to use as a guideline to improve the Web site for better trust.

We proposed an effective trust assessment framework for evaluating eCommerce Web sites based on customer preferences.

The Merriam-Webster English dictionary defines "trust" as assured reliance on the character, ability, strength, or truth of someone or something.

F. N. Egger et al [7] developed a model of trust in eCommerce called MoTEC (model of trust in eCommerce), which could classify characteristic of trust in eCommerce in terms of company, product and service, security, privacy, usability and relationship management. Our proposed framework consists of 3 factors: (1) Content, (2) design and (3) security.

3. DATA COLLECTION

This study uses one questionnaire to collect data. The most of respondents aged between 35-50 years old, while 70.2% of the respondents were male.

In the questionnaire, respondents were asked to go through the entire buying process at the e-commerce websites and do purchase the item experimentally. While they were doing this process, they had to respond to the questions based on three categories that have been defined in trust model. There were four questions on main criteria and one on trust. After the website analysis and answering the questions, lastly subjects had to identify the trust of the specific website.

Respondents that had finished analyzing of e-commerce website then were asked to rank the trust level of the e-commerce website. They had a choice of ranking the Trust level based on linguistic variable as follows: Very low, low, moderate, high and very high. It had also five options (index) ranked by 0-4 for indicating trust level as follows:

0= Very low 1=Low 2= Moderate 3= High 4= Very High

It had also 3 options (index) ranked by 0-2 for the main factors level as follows:

0= Low 1=Moderate 2= High

Respondents picked values as per their understanding of trust following the questions pertaining to trust main factors in trust model.

After ranking factors in first questionnaire, next questionnaire was published in the website and URL was provided for online respondents. They were asked to analysis three B2C e-commerce websites and then response online questionnaire. After gathering answers from SQLServer database, raw data was entered in excel for more analysis. Table 1, 2 and 3 shows frequency of 3 factors and emerged trust levels.Figure 2, 3 and 4 depict frequency of factors too.

SecurityLevel					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0.00	2	.4	.4	.4
	12.50	33	7.3	7.3	7.8
	25.00	67	14.9	14.9	22.7
	37.50	93	20.7	20.7	43.3
	50.00	83	18.4	18.4	61.8
	62.50	70	15.6	15.6	77.3
	75.00	71	15.8	15.8	93.1
	87.50	26	5.8	5.8	98.9
	100	5	1.1	1.1	100
	Total	450	100	100	

Table 1. The Frequency of Security Factor

Journal of Theoretical and Applied Information Technology
15th February 2012. Vol. 36 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

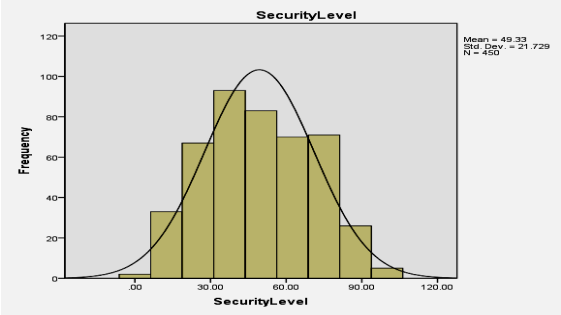

Figure 2. The Histogram of Frequency Security Factor

ContentLevel					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	12.50	25	5.6	5.6	5.6
	25.00	54	12.0	12.0	17.6
	37.50	96	21.3	21.3	38.9
	50.00	99	22.0	22.0	60.9
	62.50	92	20.4	20.4	81.3
	75.00	59	13.1	13.1	94.4
	87.50	25	5.6	5.6	100.0
	Total	450	100	100	

Table 2. The frequency of content factor

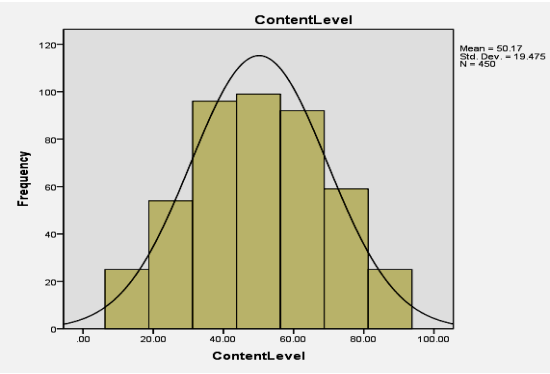

Figure 3. The Histogram of Frequency Content Factor

DesignLevel					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	3	.7	.7	.7
	12.50	34	7.6	7.6	8.2
	25.00	56	12.4	12.4	20.7
	37.50	116	25.8	25.8	46.4
	50.00	82	18.2	18.2	64.7
	62.50	81	18.0	18.0	82.7
	75.00	39	8.7	8.7	91.3
	87.50	36	8.0	8.0	99.3
	100.	3	.7	.7	100.0
	Total	450	100	100	

Table 3. The Frequency of Design Factor

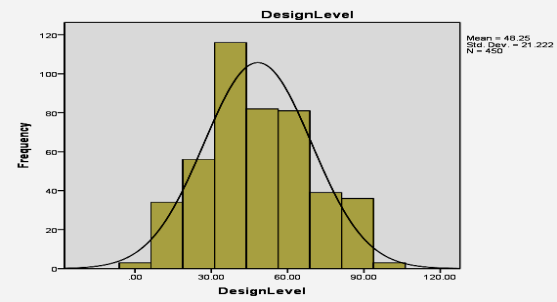

Figure 4. The Histogram of Frequency Design Factor

4. RESEARCH MODEL

The proposed model has been established based on this principle that trust in B2C websites include privacy and security, design and content.

There are some major factors in electronic commerce such as quality of websites' content. Customer attitude to uncertainty of online purchasing, affects customer's opinion. Privacy and security commitments in B2C e-commerce are reflected in the actions of the Web merchant. Yet, for consumers, the primary, visible access to privacy and security on Web merchants' sites is

Journal of Theoretical and Applied Information Technology
15th February 2012. Vol. 36 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

through statements that describe in more or less understandable terms the privacy and security policies of the Web merchant, from information collected to data sharing policies, and security features such as encryption and password protections. The privacy and security statements on today's Web sites vary from excellent and well-detailed too hard to find and difficult to read. The security systems are strongly needed to handle the process of developing the customer retention strategies in e-business transaction process in an attempt to capture the relationship within organization and with the customers. The benefits of applying trust and build up security in e-business is quite obvious.

Zang and Tarafdar [8] assumed that "content" and "design" of website are the main factors in the quality of B2C Company's website. Other experts and specialists such as Ranaganathan and Ganaouthy [9] with respect to the quality and design stated that website should contain relative information of company services. In customer's attitude, it seems that the impression of website's design is more significant than its content. Therefore, unsuited design will cause less motivation for customers to find the product or services through websites.

In this research the level of trust obtains of these three parameters performance. Figure 5 Presents a model based on our which illustrates the relationships between the different concepts.

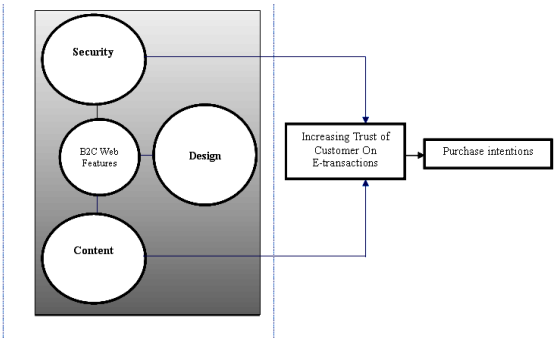

Figure 5. Proposed Trust Model

5. APPLYING K-NN METHOD FOR CLASSIFYING FACTORS

After gathering data from online customers ,k-NN method was used for classifying factors .Using k-NN method on data ,results show that which factor in which position can important on trust level . Also customers can find themselves in the situations rather than other customers.

6. MODEL VIEW

After gathering and organizing data, k-NN method was applied and desired results were obtained. Figure 6 shows a depiction of model view in predictor space. In predictor space trust levels in 5 degree has been shown. As in this depiction we can predict customer trust based on tree factors .Also in predictor space with increasing security factor, trust level to very high level has increased. in figure 6, customer with node 368 was located in very high level trust. If look at this node, we can find security level in this node is high level but content and design level are in moderate level.

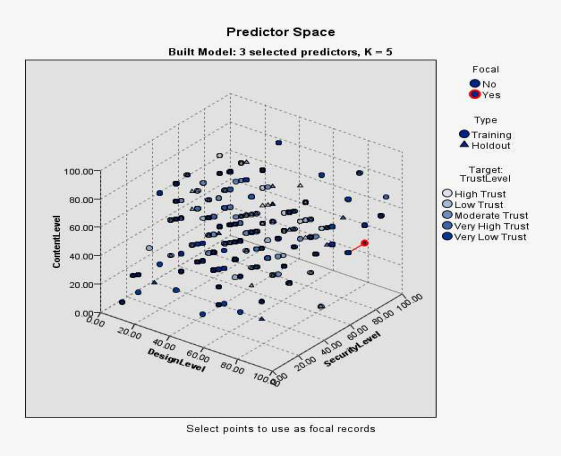

Figure 6. Nearest Neighbor Analysis Model View for Very High Level Trust

In peers chart in figure 7 we can find trust level as numeral. In this depiction also we can find that in despite of moderate level of content factor, trust remain in very high level when security is high level.

Journal of Theoretical and Applied Information Technology
15th February 2012. Vol. 36 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

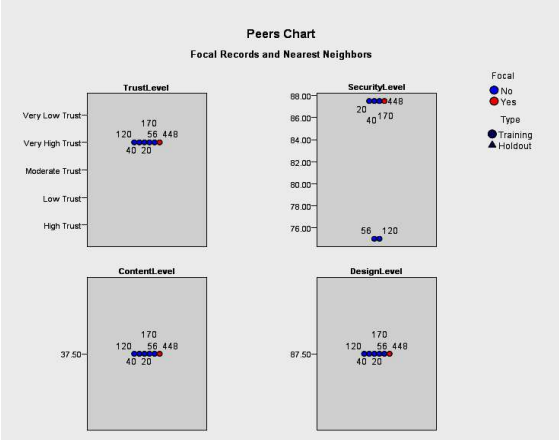

Figure 7. Peers Chart for Trust Level of Model View

In figure 8, we can find trust degree in very low level. Predictor space in this figure shows that trust level can be in very low level in despite of high level for design and content. Therefore these results indicate that security factor is very significant for trust level.

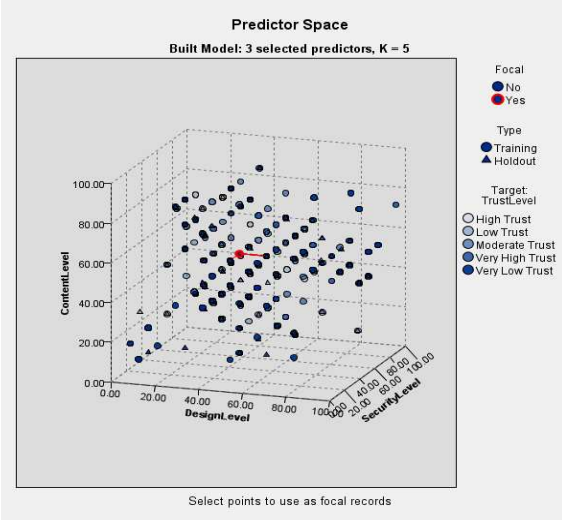

Figure 8. Nearest Neighbor Analysis Model View for Very Low Level Trust

Based on peer chart for previous figure, we can illustrate trust level as numeral for high level of design, content and low level for security. Figure 9 shows peers chart for trust level of model view.

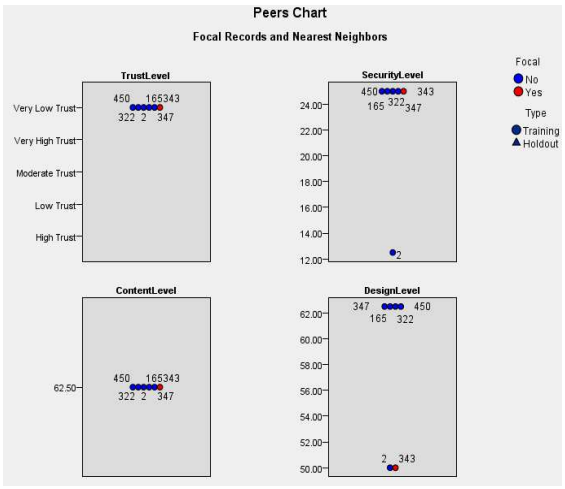

Figure 9. Peers Chart for Trust Level of Model View

7. PREDICTOR IMPORTANCE

In predictor window, we can find priority of factors generated by k-NN method. Figure 10 show that security factor is more important than two other factors. Based on predictor model and peers chart, this result had been detected. Thus figure 10 shows that predictor importance for security level is 0.43, design level is 0.30 and content level is 0.27.

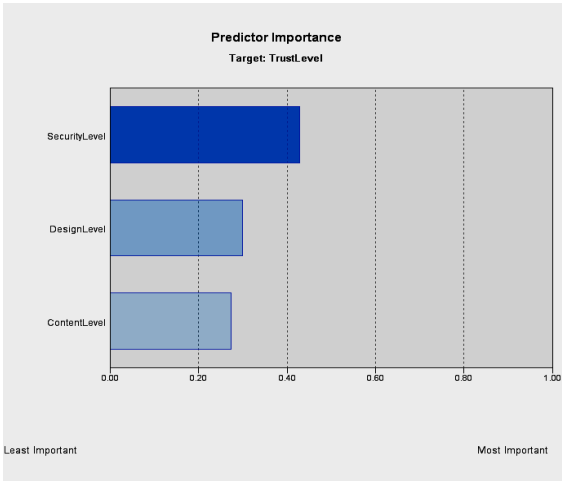

Figure 10. Predictor Importance for Comparing Security Level, Design Level and Content Level

8. SAMPLE XML OUTPUT FOR K-NN METHOD

Figure 11 shows sample exported model information to XML file.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 11. Exported Model Information to XML File

9. CONCLUSION

In this research, a survey was conducted to discover which factors on a website will influence the trust of e-commerce user most. We organized these factors as the Content, Security and Design. Using these indicators was visited existing popular websites and compared them based on a questionnaire. In this paper k-NN method was applied in the framework for analyzing trust on Web-site .Our framework consists of three factors: (1) Security, (2) Design and (3) Content. This framework gives to online customers and seller three benefits: (1) effect of factors on trust in B2C websites and percentage of them (2) recommendation for improving this web site according to security, design and content and (3) predicting new positions for locating a new customer in a B2C websites. Also the results showed that security factor is very important on trust in B2C website and predictor importance for security level is 0.43, design level is 0.30 and content level is 0.27.

REFRENCES:

[1]T.S. Bhatti, R.C. Bansal, and D.P. Kothari, "Reactive Power Control of Isolated Hybrid Power Systems", *Proceedings of International Conference on Computer Application in Electrical Engineering Recent Advances (CERA)*, Indian Institute of Technology Roorkee (India), February 21-23, 2002, pp. 626-632.

[2]B.N. Singh, Bhim Singh, Ambrish Chandra, and Kamal Al-Haddad, "Digital Implementation of an Advanced Static VAR Compensator for Voltage Profile Improvement, Power Factor Correction and Balancing of Unbalanced Reactive Loads", *Electric Power Energy Research*, Vol. 54, No. 2, 2000, pp. 101-111.

[3]J.B. Ekanayake and N. Jenkins, "A Three-Level Advanced Static VAR Compensator", *IEEE Transactions on Power Systems*, Vol. 11, No. 1, January 1996, pp. 540-545.

[1]Ripley, B.D., 1996. Pattern Recognition and Neural Networks, Cambridge University Press.

[2]Sande, I. G., "Hot-Deck Imputation Procedures", in Madow,W. G. and Olkin, I., eds., Incomplete Data in Sample Surveys, Volume3, Proceedings of the Symposium, Academic Press, 1983, pp.334-350.

[3]Chen, J. and Shao, J., "Nearest Neighbor Imputation for Survey Data", in Journal of Official Statistics, vol. 16, no. 2, 2000,pp. 113-131.

[4]Strike, K., El Emam, K. and Madhavji, N., "Software Cost Estimation with Incomplete Data", in IEEE Transactions on Software Engineering, vol. 27, 2001, pp. 890-908.

[5]Troyanskaya, O., Cantor, M., Sherlock, G., et al., "Missing Value Estimation Methods for DNA Microarrays", in Bioinformatics, vol. 17, 2001, pp. 520-525.

[6]Y. Yang (2004). "W3 Trust Model (W3TM) "A Trust-Profiling Framework to Asses Trust and Transitivity of Trust of Web-Base Service in A Heterogeneous Web Environment," Phd Thesis.

[7]F. N. Egger (2003). "Designing the Trust Experience for Business to Consumer Electronic Commerce," Ph.D. Thesis.

[8]Tarafdar M, Zhang J (2005). Analyzing the influence of Web site design parameters on Web site usability. Info. Resour. Manage. J., 18(4):62-80.

[9]Ranganathan C, Ganapathy S (2002). Key dimensions of business-toconsumer Web sites. Info. Manage., 39(6): 457-465.

[10] Athitsos, V., Alon, J., Sclaroff, S.: Efficient nearest neighbor classification using a cascade of approximate similarity measures. In: CVPR '05, pp. 486–493. IEEE

Computer Society, Washington, DC, USA (2005)

[11]Athitsos, V., Sclaroff, S.: Boosting nearest neighbor classifiers for multiclass recognition.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In: CVPR '05, IEEE Computer Society, Washington, DC, USA (2005)

[12] Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1), 21–27 (1967)

[13] Zhang, H., Berg, A.C., Maire, M., Svm-knn, J.M.: Discriminative nearest neighbor classification for visual category recognition. In: CVPR '06, pp. 2126–2136. IEEE Computer Society, Los Alamitos, CA, USA (2006)

[14]www.scss.tcd.ie/~jzhou./tutorial10/tutorial-10-K%20nearest%20neighbours.doc.

[15] Peng, J., Heisterkamp, D.R., Dai, H.K.: LDA/SVM driven nearest neighbor classification. In: CVPR '01, p. 58. IEEE Computer Society, Los Alamitos, CA, USA (2001)