The Nucleus **50**, No. 4 (2013) 387-394

The Nucleus
A Quarterly Scientific Journal of Pakistan Atomic Energy Commission
NCLEAM, ISSN 0029-5698

EVALUATION OF SIMILARITY MEASURES FOR CATEGORICAL DATA

T. HUSSAIN and *S. ASGHAR[1]

Department of Computer Sciences, Muhammad Ali Jinnah University (MAJU), Islamabad, Pakistan

[1] Institute of Information Technology, University of Arid Agriculture, Rawalpindi, Pakistan

(Received June 16, 2013 and accepted in revised form November 27, 2013)

Similarity among the objects is a fundamental concept to almost all the technical field such as *information retrieval; data mining; mathematics;* and *bioinformatics.* A similarity measure symbolizes relation among the objects, which can be either, documents, queries or features of any database. Similarity measure helps to rank the objects in accordance to their importance in specific data mining application. A similarity measure is a function that computes the degree of similarity between a pair of objects. Similarity base applications are countless. Data mining is used to build the knowledge base of the large data repositories for human inferences and analysis. Data mining techniques are more frequent in all such technical fields where the similarity as' required. The proper selection of similarity or distance measure is a key to many data mining techniques such Clustering; Classification; and Outlier Detection. For categorical data, computation of similarity measure is a complex phenomenon. The measures used for continues data such as Euclidean Measures are generalized upto some extent and can be applied in any continues data domain. Euclidean measures are widely applied to categorical data without considering the domain knowledge and nature of categorical data. Due to the complex nature of categorical data, no standard measure like Euclidean is available in literature. In this paper, we are evaluating the different categorical measures in accordance with their usage in different data mining applications and techniques. We are also proposing the chi-fuzzy measure to address the categorical data issue.

Keywords : Categorical data, Similarity measure

1. Introduction

A similarity measure symbolizes relation among the objects, which can be either, documents, queries or features of any database. Similarity measure helps to rank the objects in accordance to their importance in specific data mining application. A similarity measure is defined as a function that computes the degree of similarity between a pair of objects [1]. Similarity or distance measure between two objects or entities plays core role in data mining applications for knowledge discovery where the objects have to be classified on the basis of distance computations. Data mining applications such as Clustering; Classification and distance based Outlier Detection require the similarity or distance measure between their objects [2-4]. Generally, these applications can be divided into three steps as shown in Figure 1. These steps are: *data pre-processing*; applying *similarity measure* and *classifier approach*. If we are able to find out how much similar are the data objects, we can have better results of classifier. Similarity measure gives us the precision and accuracy of closeness of relationship between objects. In the eyes of [5], similarity measure plays more significant role than classifier. It can be apprehended that proper selection of similarity measure is key process.

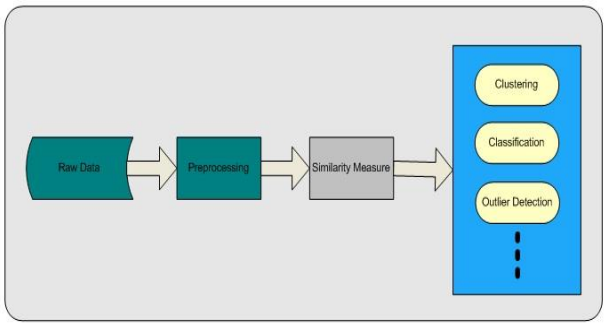

Figure 1. Steps of similarity measure based approaches.

The selection of similarity measure is also dependent on the data type as well. For continues data, *Euclidean Family Measures* are widely used to compute the distance between objects. The most popular distance measures of this family are: *Euclidean; Minkowski; Manhatton; Angular Separation* and *Cnaberra Distances* [6]. The computation of similarity between categorical data set is a complex phenomena. Categorical data take up on the different categorical values, so we cannot compare them directly [7, 8].

Categorical data item set can take different values for their categorical attributes and these item values cannot be set in order [9, 10]. Table 1 explains the categorical data. Sex can take either

* Corresponding author : Sohail.Asg@gmail.com

values of male or female, but we cannot measure distance between male and female directly. Similarly, the build can be low, medium or high. We cannot apply traditional distance measures (Euclidean Family Measures) to compute the distance between such type of data.

Table 1. Categorical data.

Sex	Male, Female
Build	Low, Medium, High
Complexion	Fair, Wheatish, Whitish

For such type of categorical data, we require a similarity measure which correlates the attributes values in the true sense of categorical data. Categorical data are complex data structure, consequently, similarity computation as also a complex one. We may require to study the different characteristics of categorical data such as data size; number of categorical attributes; nature of categorical data attributes with respect to values taken by these attributes; and frequency distribution of categorical values in different attributes [1].

Numbers of categorical measures are available in the literature, but the proper selection of similarity measure is guarantee of proper and to the point results of classifier for predefined data mining technique. Different categorical measures have been evaluated from data driven point of view [7]. His research work also brought the fourteen categorical measures in a single context. There are number of researches in literature, which are evaluating categorical measure in different dimensions. It has been observed that we lack hierarchical taxonomy of the categorical measures in the literature. There are multiple advantages of taxonomy. One of the biggest advantage is the proper selection of categorical measure as per data and domain requirement.

1.1 Research Contributions

Followings are the research contributions or research objectives addressed in this paper.

- To study the different categorical measures in the prospective of their applications in data mining. How effective, a similarity measure can be in the different techniques? Categorization will be helpful in many dimensions.
- Classified similarity measures will be evaluated on the data set to ensure the utility of measures.
- Proposed categorical measure based on chi-square and fuzzy logic.

1.2 Paper Structure

The rest of paper is being structured as follows: Section 2 elaborates the detailed literature review on exiting categorical similarity measures. While in section 3, we discussed in detail the categorical data. In section 4, similarity measure and its importance has been highlighted. In section 5, we proposed a novel the categorical similarity measure along with results and in section 6, we discussed the future research directions.

2. Literature Review

The similarity measure for categorical data has always been complex and laborious task. For continuous data, Euclidean, Minkiwski, Manhattan distances are being renowned measures, although there exists vast variety of measures and selection of distance metric is easy as compare to categorical data.

In the early seventies, the main concern of scientists and researchers were the biological data. In their research Sneath and Sokal [11], produced the comprehensive taxonomy of similarity measures for biological and ecology. Their research contribution was the cornerstone to address the categorical data and still used as beam research. “Simple matching similarity” is one of the fundamental measures for categorical data sets.

Goodall [12] developed the probability based similarity measure and addressed the measure in various attributes types and characteristics of attributes such as Qualitative Attributes; Ordered Attributes; and Metrical Attributes. Eq. (1) explains the similarity measure that it assigns higher similarity if i = j for matching attributes.

$$S_{ij} = 1 - \sum_{k \in Q} \frac{f_k(f_k - 1)}{m(m-1)}, i = j$$
$$S_{ij} = 0, i \neq j \quad (1)$$

“A comparative evaluation of categorical data” has been produced [7]. Their valuable contribution is to evaluate the different measures on different data set. A taxonomy for categorical measures [7] and classify the measures into three broader categories such as *Diagonal Entries Only Measures*; *Off Diagonal Entries*; and both *Diagonal and Off Diagonal Entries.* There are three new categorical measures Goodall2, Goodall3 and

Goodall4 produced [7] based on Goodall measure [12]. Their proposed measures are claimed to be based data driven model. Goodall2 assigns higher similarity value if the matching values are infrequent.

Ahmad and Dey [13] proposed the measure by assigning the distance to attributes based on effect of attribute on dataset. The distance was calculated as in Eq. (2)

$$\delta(x,y) = sum / (1-m) \tag{2}$$

Where x and y are two defined attributes of a given dataset and there are *m* number of total attributes in the dataset. A similarity measure based on cardinality of domain of attribute values of two objects is proposed in [9]. For matching attributes, distance is calculated as 1 and for mismatching values 0. Eq. (3) gives details about the calculation of similarity between two objects.

$$sim(O_i,O_j) = 1 - \sum_{h=1}^{m} d(x_h,y_h) / \sum_{k \in x_h = y_h} D_k \tag{3}$$

Le and Ho [14] applied the phenomena of conditional probability distribution cpd to calculate the similarity of two values of an attribute. Eq. (4) calculates the similarity of attribute A_i.

$$\varphi(v_i,v_i^{'}) = \sum_{j,j \neq i} \psi(cpd(A_j \mid A_i = v_i), cpd(A_j \mid A_i = v_i^{'})) \tag{4}$$

There are a number of ways to calculate similarity measure between categorical attributes/objects but binary-based similarity measures are the pioneer for categorical data [14, 15]. For the first time Sneath and Sokal [11] in their book on numerical taxonomy discussed the number of binary-based measures that are commonly used for categorical data. [15] reviewed the most commonly used binary measures such as Simple Matching [11]; Tanimioto (Jaccard); and Hamming[6]. In Simple Matching, only matched attribute values are picked in two attributes as explained in following Eq. (5).

$$S_M(X,Y) = n(X \cap Y) \tag{5}$$

For simple matching measure for categorical data the size of data set is not considered.In Tanimioto measure is composed of ratio common elements to the number of all different elements in the two objects. This measure is also known as Jaccard Measure and given in Eq. (6).

$$S_T(A,B) = \frac{n(A \cap B)}{n(A \cup B)} = \frac{n(A \cap B)}{n(A) + n(B) - n(A \cap B)} \tag{6}$$

The literature review can be summarized according to [7] that there is no categorical similarity measure which is performing best in all situations. Moreover, it can be concluded that domain, data mining technique and characteristics of categorical data are to analyzed carefully before applying and selection of categorical measure Table 2.

3. Categorical Data

For the last two decades, trends in data retrieving, storing, manipulating and exploiting have been changed. There are varieties of data structures available. Random variables are divided into two types and these types are: numerical data and categorical data. We can say that two different types of data variables (categorical and numerical variables) produce the two different types of data. Categorical variable produce data into different categories and numerical variables produce data in numerical form. We can differentiate between categorical variables and numerical variable by a simple method of responses of these variables. Responses to such questions as "What is your favourite color?" Response is red; blue; green; or yellow. “What is your sex?” Male or female. These reposes are leading or differentiating these variables as categorical variables. Categorical data is also known as nominal data or qualitative data. Categorical data is not in any proper order.

While the responses to such a question as “How many employees are there in your company? The answer may be 60 or 100 or more. Similarly, in response to these questions as in numerical such as height and weight of a person. Numerical data can be either discrete or continuous. Table 3 below may help us more clearly to visualize the differences between these two variables. Numerical data can also be subdivided into two subcategories such as Discrete and Continuous data types.

Categorical data format is one of the fundamental data type in Computer science. The calculation of similarity measure for categorical attributes or objects are even a tricky mission. Similarity measure should preserve the essence of categorical data.

3.1 Characteristics of Categorical Data

The characteristics of data type plays prominent role to identify the true behavior of data set [16].

Table 2. Measure evaluation.

	Author	Measure	Similarity	Range	Application Area	Technique	Comparing Measure
Overlapping Measures	Hamming (1950)	Hamming Distance	$d_H(x,y)=\sum_{i=1}^{n}\delta(x_i,y_i)$	(0,1)	Clustering	SOM	Euclidean
	Lourenco (2004)	Jaccard	$S_T(A,B)=\frac{n(A\cap B)}{n(A\cup B)}=\frac{n(A\cap B)}{n(A)+n(B)-n(A\cap B)}$	(0,1)	Clustering	SOM	Anderberg
	Sokal-Michener	Simple Matching	$S_M(A,B)=n(A\cap B)$	(0,1)			
Probabilistic Model	Eskin (2002)	Eskin	$\begin{cases}0\\ \frac{n_k^2}{n_k^2+2}\end{cases}$	(2/3, $n^2/(n^2+2)$)	Outlier Detection	kNN	Boriah (2008)
	Anderberg (1973)	Anderberg	a /a + 2(a+b)	(0,1)			
	Boriah (2008)	Goodall2	$1-\sum p_k^2(q)(if(x_k=y_k))$	(0,1 – 2/N(N-1))	Outlier Detection	NN	Boriah (2008)
	Boriah (2008)	Goodall3	$1-p_k^2(x_k)if(x_k=y_k)$	(0,1 – 2/N(N-1))	Outlier Detection	NN	Boriah (2008)
	Boriah (2008)	Goodall4	1 – Goodall3	(2/N(N-1),1)	Outlier Detection	NN	Boriah (2008)
	Le (2005)	Le-Ho	$\varphi_{Ai}(v_i,v_i')=\sum_{j,j\neq i}\phi(cpd(A_j/A_i=v_i),cpd(A_j/A_i=v_i'))$		Classification	NN	Goodall Gower-Legendre
	Goodall (1966)	Goodall	$S_{ij}=1-\sum_{k\in Q}\frac{f_k(f_k-1)}{m(m-1)}, i=j$ $S_{ij}=0, i\neq j$	(0,1)	Bio		Sorensen
Data Driven Model	Amir – Lipika (2006)	Amir – Lipika (2006)	$\delta(x,y)=1/(m-1)(\sum_{j=1}^{m}\delta(x,y,A_i)$		Clustering	K-Modes	
	Boriah (2008)	Lin1	$\sum_{q\in Q}\log P_k(q)$	(-2log N, 0)	Outlier Deduction	kNN	Boriah (2008)
	Aranganayagi (2009)	Aranganayagi (2009)	$1-\sum_{h=1}^{m}d(x_h,y_h)/\sum_{k\in(x_h=y_h)}D_k$	(0,1)	Clustering	K-Modes	Simple Matching

Before calculating the similarity measure, following characteristics of categorical data should be known [7, 15]. Order of Categorical data: Categorical data has no single order. It can be ordered in several ways. Visualization: Categorical data can be visualized in specific order in particular domain. Structure: Categorical data has no well defined structure.

Attributes: Data mining techniques and performance of these techniques is directly proportional to number of categorical attributes in dataset. Data set can be multivariate and multi attribute. Occurrence (frequency) of particular attributes and size of data set are also very important for the calculation of similarity for categorical data.

Table 3. Data types.

Model Questions	Response	Data Type
What is your favorite color?	Red, Blue, Green, Yellow	Categorical
Sex of individual?	Male/ Female	Categorical
What is your weight?	67 kg	Numerical
How old you are?	35 years	Numerical

4. Similarity / Dissimilarity Measure

Similarity measure for categorical data or dissimilarity measure, convey the similar meaning in either way. Both are reciprocal to each other. For any two categorical attributes x_i and y_i if the similarity $\delta(x_i, y_i) = 1$ then dissimilarity will be 0 and vice versa. Similarity between categorical attributes shows the strength of bond between them. We can infer how strongly the two categorical attributes are related to each other. and for dissimilarity, we calculate that how dissimilar two categorical objects.

While calculating the distance between two object is find out how apart they are. In conventional methods, we apply the standard distance metric such as Euclidean, Manhattan or Minkiwiski metrics to categorical data objects and it can be converted to similarity. The distance between two objects x_i and y_i is computed as $D(x_i, y_i)$ and similarity can be computed as [17] Eq. 7.

$$\delta(x_i, y_i) = \frac{1}{1 + D(x_i, y_i)} \quad (7)$$

Similarity of categorical data posses the same properties which are applicable and defined for distance metric. Following properties should hold for similarity measures [15, 16].

1. $0 \le \delta(x_i, y_i) \le 1$
2. $\delta(x_i, x_i) = 0$
3. $\delta(x_i, x_i) = \delta(y_i, x_i)$

4.1 Example

There are several categorical measures have been proposed in literature. Some measures show better results in classification, some measures have performed well in clusters and while some are very outstanding in outlier detection as shown in Table 1 [18]. Following example has been inspired by the similarity measure proposed by Le and Ho [14]. Similarity (dissimilarity) has been calculated through following Eq. 8.

$$\varphi(v_i, v_i^{'}) = \sum_{j, j \neq i} \psi(cpd(A_j \mid A_i = v_i), cpd(A_j \mid A_i = v_i^{'})) \quad (8)$$

Table 4, Example data.

	Data set values		Probabilities		
Sex / Complexion	Male	Female	P (Male)	P (Female)	Sum
Fair	15	20	0.42857	0.57142	1
Wheatish	25	15	0.625	0.375	1
Whitish	10	15	0.4	0.6	1

The similarities between the objects (Fair, Wheatish), (Fair, Whitish) and (Wheatish, Whitish) has been calculated as:

$$\delta(Fair, Wheatish) = 0.226286$$

$$\delta(Fair, Whitish) = 0.004855$$

$$\delta(Wheatish, Whitish) = 0.297434$$

In the following Figure 2 we have plotted the graph between attributes of complexion and their distribution in dataset.

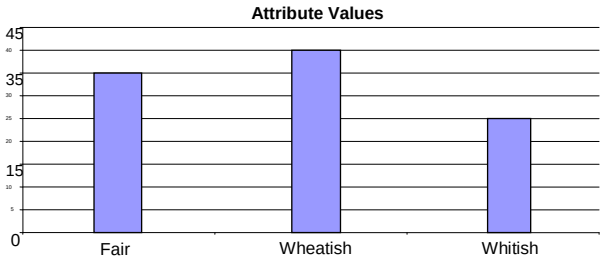

Figure 2. Attributes distribution.

Figure 3 shows the similarities of an attribute of the categorical objects.

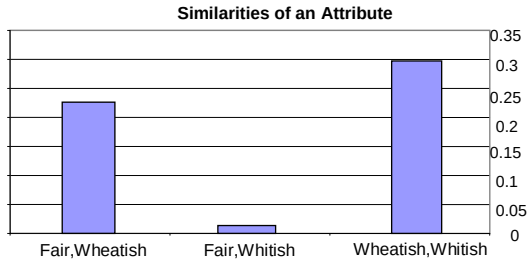

Figure 3. Categorical similarities of an attribute.

5. Proposed Categorical Measure

There are a number of statistical techniques, which have been applied to categorical data sets to find out the similarity between the categorical variables. In statistics, one of the oldest and commonly used tests to find out the association between two variables is the Chi-Square test. The Karl Pearson's Chi-Square test is an exceptionally powerful test to find out the nonparametric association between two categorical variables. Nonparametric tests are applied when the data are not normally distributed. This test is most appropriate when data are in the form of frequencies.

As we know that categorical and numerical data are two fundamental data types of a random variable. There are a number of measures available to test these variables. One of them is a chi square (X^2) measure, which is widely used to examine the categorical variables in different domains. As we know that categorical variable originate data into different categories accordingly while numerical, variables produce the data in numerical form as well. Most of the time we depend on the answers of different queries such as "What is your favorite color?" The answer might be red/green/blue then it represents the categorical variables and categorical data as well. In the same way, answer the question such as "What is your typing speed?" The response will be in numerical form then these variables represent the numerical data. Furthermore, the numerical data can be divided into two subcategories such as discrete and continuous data.

The concept of Fuzzy Logic (FL) is commonly used in the computer science field [19] where data processing is done in the form partial membership instead of crisp data or non-membership[10]. FL incorporates a simple, rule-based solution to the problems such as if X and Y then Z. These techniques are most helpful in solving the control problem rather than modeling a system mathematically. The FL models are pragmatic and relying on an operator's expertise and domain knowledge rather than technicality of the system.

The use of fuzzy logic in complex problem solving is the cornerstone because FL is naturally vigorous and dynamic since it does not require precise, accurate and noise-free inputs and can be managed programmatically to failures in safely due to feedback sensor stops working [20]. The output control generated through FL is a smooth control rather than a wide range of parameters. Since the FL controllers are controlled manually, this can improve overall performance and reliability of the system.

For these reasons, we apply Chi-Fuzzy to compute the similarity between the categorical objects. In following the steps we apply the chi-square and fuzzy logics to the data set. Table 4 shows the categorical variables and their respective values.

Figure 4 shows the disperness of categorical variables while Figure 5 shows the graph of categorical variable based on their values. In Table 5, we have calculated the chi square values of the variables.

Table 5. Categorical data objects and values.

	Red	**Green**	**Blue**	
Ö	50	40	10	100
◊	21	21	28	70
Δ	24	24	72	120
	95	85	110	290

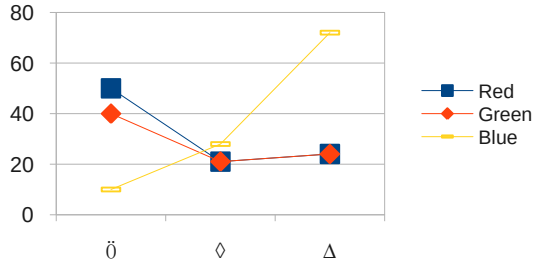

Figure 4. Dispersness of categorical values.

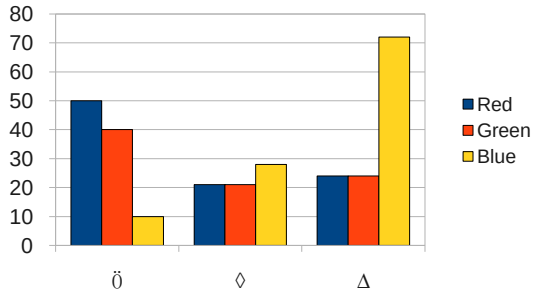

Figure 5. Graph of categorical values.

In the next steps we are calculating the chi-square value for each pair of categorical variable.

Table 6. Chi Square values of variables.

	Red	Green	Blue	
ö	50	40	10	100
◊	21	21	28	70
	71	61	38	170
χ^2	21.67			
	Red	Green	Blue	
◊	21	21	28	70
Δ	24	24	72	120
	45	45	100	190
χ^2	7.09			
	Red	Green	Blue	
ö	50	40	10	100
Δ	24	24	72	120
	74	64	82	220
χ^2	58.67			

After calculating the values of χ^2 of each pair of categorical variables, we calculated the critical region (Eq. 9) value by using χ^2 and chi square table as

$$\chi^2(0.5)(2) \quad (9)$$

5.1. Chi-square Test of Independence

Now we apply the chi square test. With the help this test, we will monitor the dependency of corresponding variables. In a test of independence, two hypotheses are defined. These hypotheses measure the state of respective variables. The hypotheses can be defined as below:

Ho: The two categorical variables are independent.

Ha: The two categorical variables are to relate.

In Figure 6 we have shown the graph of Contingency values of Categorical variables based on the chi square.

5.2. Fuzzification Step

Now in this last step we fuzzify the Chi-square values of the variables in following Table 7. We divided the values into three fuzzy regions such as *Least Similarity*; *Average Similarity* and *Absolute Similarity*.

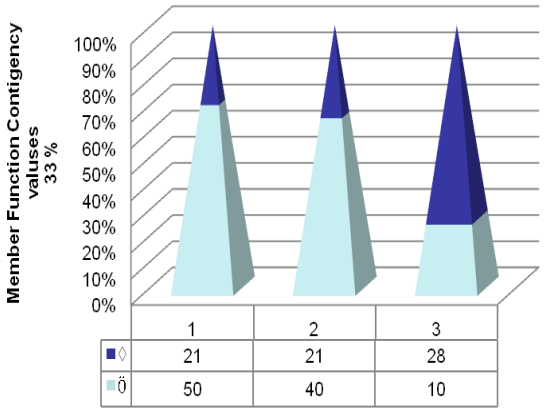

Figure 6. Contingency values of categorical variables.

Table 7. Fuzzification of categorical variables.

		Fuzzification on Chi Values		
	χ^2	Least Similarity	Average Similarity	Absolute Similarity
(ö,◊)	21.67	0	0.9	0
(◊,Δ)	7.09	0.8	0	0
(ö,Δ)	58.68	0	0	1

6. Conclusion

Similarity measure for categorical data is complex with respect to technique applied, categorical data format and domain. There is no single similarity measure which fits in all scenarios. We have to be very specific about the selection of similarity measure. A similarity measure may perform well for classification and may not provide optimum results in clustering and outlier detection. For this reason the selection and calculation of similarity measure for categorical data is challenging task.

For future work, we will test the some of the renowned categorical measures in the context of different data mining techniques and we may be able to categorize the best categorical measure in respective technique of data mining. We will also propose new similarity measure which may suit in all flavors and environment.

References

[1] M.Y. Shih, J.W. Jheng and L. F. Lai, Tamkang Journal of Science and Engineering **13**, No. 1 (2010) 11.

[2] G. Fung, A Comprehensive Overview of Basic Clustering Algorithms, www.cs.wisc.edu/~gfung/clustering.ps.gz. (June 22, 2001).

[3] H. Lu and T.T.S. Nguyen, Experimental Investigation of PSO Based Web User Session Clustering, International Conference of Soft Computing and Pattern Recognition, IEEE (2009).

[4] Z. Ma and O.R.L. Sheng, Clustering Web Session Using Extended General Pages, Proceedings of 8th Pacific Asia Conference on Information Systems, Shangia, China (2004) p. 5.

[5] L. Chaofeng, Research on Web Session Clustering, Journal of Software **4**, No. 5 (2009) 460.

[6] T. Hussain, S. Asghar, and S. Fong, A Hierarchical Cluster Based Preprocessing Methodology for Web Usage Mining. 6th International Conference on Advanced Information Management and Service (IMS). Seoul, Korea (2010).

[7] S. Boriah, V. Chandola and V. Kumar, Similarity Measure for Categorical Data: A Comparative Evaluation, Proceedings of the Eighth SIAM International Conference on Data Mining (2008).

[8] C.M. Nichele and K. Becker, Clustering Web Sessions by Levels of Page Similarity Springer-Verlag Berlin Heidelberg (2006) pp. 346-350.

[9] S. Aranganayagi, K. Thangavel and S. Sujatha, New Distance Measure based on the Domain for Categorical Data. ICAC, IEEE (2009).

[10] Z.C. Johanyak and S. Kovacs, Distance Based Similarity Measure of Fuzzy Sets (2004).

[11] P.H.A. Sneath and R.R. Sokal, Numerical Taxonomy: The Principles and Practice of Numerical Classification, San Francisco: W. H. Freeman and Company (1973).

[12] D.W. Goodall, Biometrics, **22**, No. 4 (1966) 882.

[13] A. Ahmad and A. Dey, ScienceDirect, Pattern Recognition Letters **28** (2006) 110.

[14] S.Q. Le and T.B. Ho, Elsevier **26** (2005) 2549.

[15] F. Lourenco, V. Lobo and F. Bacao, Binary-Based Similarity Measures for Categorical Data and Their Application in Self Organizing Maps, JOCLAD 2004 - XI Jornadas de Classificacao e Anlise de Dados, Lisbon, April 1-3, (2004).

[16] V. Chandola, S. Boriah and V. Kumar, A Framework for Exploring Categorical Data, SIAM (2009) pp.187-198.

[17] M. Setnes, R. Babuˇska, U. Kaymak and H.R.V.N Lemke, Cybernetics **28**, No. 3 (1998) 376.

[18] W. Wang and O.R. Zaiane, Clustering Web Sessions by Sequence Alignment, Third International Workshop on Management of Information on the Web in Conjunction with 13th International Conference on Database and Expert Systems Applications (2002) pp. 394–398.

[19] A. Ahmad and L. Dey, Algorithm for Fuzzy Clustering of Mixed Data with Numeric and Categorical Attributes. Springer -Verlag Berlin Heidelberg (2005) pp. 561 – 572.

[20] G. Castellano, F. Mesto, M. Minunno and M. Torsello, A. Web User Profiling Using Fuzzy Clustering, Springer (2007) pp. 94-101.