A Survey of Binary Similarity and Distance Measures

Seung-Seok Choi, Sung-Hyuk Cha, Charles C. Tappert
Department of Computer Science, Pace University
New York, US

ABSTRACT

The binary feature vector is one of the most common representations of patterns and measuring similarity and distance measures play a critical role in many problems such as clustering, classification, etc. Ever since *Jaccard* proposed a similarity measure to classify ecological species in 1901, numerous binary similarity and distance measures have been proposed in various fields. Applying appropriate measures results in more accurate data analysis. Notwithstanding, few comprehensive surveys on binary measures have been conducted. Hence we collected 76 binary similarity and distance measures used over the last century and reveal their correlations through the hierarchical clustering technique.

Keywords: binary similarity measure, binary distance measure, hierarchical clustering, classification, operational taxonomic unit

1. INTRODUCTION

The binary similarity and dissimilarity (distance) measures play a critical role in pattern analysis problems such as classification, clustering, etc. Since the performance relies on the choice of an appropriate measure, many researchers have taken elaborate efforts to find the most meaningful binary similarity and distance measures over a hundred years. Numerous binary similarity measures and distance measures have been proposed in various fields.

For example, the *Jaccard* similarity measure was used for clustering ecological species [20], and *Forbes* proposed a coefficient for clustering ecologically related species [13, 14]. The binary similarity measures were subsequently applied in biology [19, 23], ethnology [8], taxonomy [27], image retrieval [25], geology [24], and chemistry [29]. Recently, they have been actively used to solve the identification problems in biometrics such as fingerprint [30], iris images [4], and handwritten character recognition [2, 3]. Many papers [7, 16, 17, 18, 19, 22, 26] discuss their properties and features.

Even though numerous binary similarity measures have been described in the literature, only a few comparative studies collected the wide variety of binary similarity measures [4, 5, 19, 21, 28, 30, 31]. Hubalek collected 43 similarity measures, and 20 of them were used for cluster analysis on fungi data to produce five clusters of related coefficients [19]. Jackson et al. compared eight binary similarity measures to choose the best measure for ecological 25 fish species [21]. Tubbs summarized seven conventional similarity measures to solve the template matching problem [28], and Zhang et al. compared those seven measures to show the recognition capability in handwriting identification [31]. Willett evaluated 13 similarity measures for binary fingerprint code [30]. Cha et al. proposed weighted binary measurement to improve classification performance based on the comparative study [4].

Few studies, however, have enumerated or grouped the existing binary measures. The number of similarity or dissimilarity measures was often limited to those provided from several commercial statistical cluster analysis tools. We collected and analyzed 76 binary similarity and distance measures used over the last century, providing the most extensive survey on these measures.

This paper is organized as follows. Section 2 describes the definitions of 76 binary similarity and dissimilarity measures. Section 3 discusses the grouping of those measures using hierarchical clustering. Section 4 concludes this work.

2. DEFINITIONS

Table 1 OTUs Expression of Binary Instances *i* and *j*

j \ *i*	1 (Presence)	0 (Absence)	Sum
1 (Presence)	$a = i \bullet j$	$b = \bar{i} \bullet j$	$a+b$
0 (Absence)	$c = i \bullet \bar{j}$	$d = \bar{i} \bullet \bar{j}$	$c+d$
Sum	$a+c$	$b+d$	$n=a+b+c+d$

Suppose that two objects or patterns, i and j are represented by the binary feature vector form. Let n be the number of features (attributes) or dimension of the feature vector. Definitions of binary similarity and distance measures are expressed by *Operational Taxonomic Units* (OTUs as shown in Table 1) [9] in a 2 x 2 contingency table where a is the number of features where the values of i and j are both 1 (or presence), meaning '*positive matches*', b is the number of attributes where the value of i and j is (0,1), meaning 'i absence *mismatches*', c is the number of attributes where the value of i and j is (1,0), meaning 'j absence mismatches', and d is the number of attributes where both i and j have 0 (or absence), meaning '*negative matches*'. The diagonal sum $a+d$ represents the total number of matches between

i and *j*, the other diagonal sum *b*+*c* represents the total number of mismatches between *i* and *j*. The total sum of the 2x2 table, *a*+*b*+*c*+*d* is always equal to *n*.

Table 2 [5] lists definitions of 76 binary similarity and distance measures used over the last century where *S* and *D* are similarity and distance measures, respectively.

Table 2 Definitions of Measures for binary data

$$S_{JACCARD} = \frac{a}{a+b+c} \quad (1)$$

$$S_{DICE} = \frac{2a}{2a+b+c} \quad (2)$$

$$S_{CZEKANOWSKI} = \frac{2a}{2a+b+c} \quad (3)$$

$$S_{3W-JACCARD} = \frac{3a}{3a+b+c} \quad (4)$$

$$S_{NEI\&LI} = \frac{2a}{(a+b)+(a+c)} \quad (5)$$

$$S_{SOKAL\&SNEATH-I} = \frac{a}{a+2b+2c} \quad (6)$$

$$S_{SOKAL\&MICHENER} = \frac{a+d}{a+b+c+d} \quad (7)$$

$$S_{SOKAL\&SNEATH-II} = \frac{2(a+d)}{2a+b+c+2d} \quad (8)$$

$$S_{ROGER\&TANIMOTO} = \frac{a+d}{a+2(b+c)+d} \quad (9)$$

$$S_{FAITH} = \frac{a+0.5d}{a+b+c+d} \quad (10)$$

$$S_{GOWER\&LEGENDRE} = \frac{a+d}{a+0.5(b+c)+d} \quad (11)$$

$$S_{INTERSECTION} = a \quad (12)$$

$$S_{INNERPRODUCT} = a+d \quad (13)$$

$$S_{RUSSELL\&RAO} = \frac{a}{a+b+c+d} \quad (14)$$

$$D_{HAMMING} = b+c \quad (15)$$

$$D_{EUCLID} = \sqrt{b+c} \quad (16)$$

$$D_{SQUARED-EUCLID} = \sqrt{(b+c)^2} \quad (17)$$

$$D_{CANBERRA} = (b+c)^{\frac{2}{2}} \quad (18)$$

$$D_{MANHATTAN} = b+c \quad (19)$$

$$D_{MEAN-MANHATTAN} = \frac{b+c}{a+b+c+d} \quad (20)$$

$$D_{CITYBLOCK} = b+c \quad (21)$$

$$D_{MINKOWSKI} = (b+c)^{\frac{1}{1}} \quad (22)$$

$$D_{VARI} = \frac{(b+c)}{4(a+b+c+d)} \quad (23)$$

$$D_{SIZEDIFFERENCE} = \frac{(b+c)^2}{(a+b+c+d)^2} \quad (24)$$

$$D_{SHAPEDIFFERENCE} = \frac{n(b+c)-(b-c)^2}{(a+b+c+d)^2} \quad (25)$$

$$D_{PATTERNDIFFERENCE} = \frac{4bc}{(a+b+c+d)^2} \quad (26)$$

$$D_{LANCE\&WILLIAMS} = \frac{b+c}{(2a+b+c)} \quad (27)$$

$$D_{BRAY\&CURTIS} = \frac{b+c}{(2a+b+c)} \quad (28)$$

$$D_{HELLINGER} = 2\sqrt{\left(1-\frac{a}{\sqrt{(a+b)(a+c)}}\right)} \quad (29)$$

$$D_{CHORD} = \sqrt{2\left(1-\frac{a}{\sqrt{(a+b)(a+c)}}\right)} \quad (30)$$

$$S_{COSINE} = \frac{a}{\sqrt{(a+b)(a+c)^2}} \quad (31)$$

$$S_{GILBERT\&WELLS} = \log a - \log n - \log(\frac{a+b}{n}) - \log(\frac{a+c}{n}) \quad (32)$$

$$S_{OCHIAI-I} = \frac{a}{\sqrt{(a+b)(a+c)}} \quad (33)$$

$$S_{FORBESI} = \frac{na}{(a+b)(a+c)} \quad (34)$$

$$S_{FOSSUM} = \frac{n(a-0.5)^2}{(a+b)(a+c)} \quad (35)$$

$$S_{SORGENFREI} = \frac{a^2}{(a+b)(a+c)} \quad (36)$$

$$S_{MOUNTFORD} = \frac{a}{0.5(ab+ac)+bc} \quad (37)$$

$$S_{OTSUKA} = \frac{a}{((a+b)(a+c))^{0.5}} \quad (38)$$

$$S_{MCCONNAUGHEY} = \frac{a^2-bc}{(a+b)(a+c)} \quad (39)$$

$$S_{TARWID} = \frac{na-(a+b)(a+c)}{na+(a+b)(a+c)} \quad (40)$$

$$S_{KULCZYNSKI-II} = \frac{\frac{a}{2}(2a+b+c)}{(a+b)(a+c)} \quad (41)$$

$$S_{DRIVER\&KROEBER} = \frac{a}{2}(\frac{1}{a+b}+\frac{1}{a+c}) \quad (42)$$

$$S_{JOHNSON} = \frac{a}{a+b}+\frac{a}{a+c} \quad (43)$$

$$S_{DENNIS} = \frac{ad-bc}{\sqrt{n(a+b)(a+c)}} \quad (44)$$

$$S_{SIMPSON} = \frac{a}{\min(a+b,a+c)} \quad (45)$$

$$S_{BRAUN\&BANQUET} = \frac{a}{\max(a+b,a+c)} \quad (46)$$

$$S_{FAGER\&McGOWAN} = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{\max(a+b,a+c)}{2} \quad (47)$$

$$S_{FORBES-II} = \frac{na-(a+b)(a+c)}{n\min(a+b,a+c)-(a+b)(a+c)} \quad (48)$$

$$S_{SOKAL\&SNEATH-IV} = \frac{\frac{a}{(a+b)}+\frac{a}{(a+c)}+\frac{d}{(b+d)}+\frac{d}{(b+d)}}{4} \quad (49)$$

$$S_{GOWER} = \frac{a+d}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad (50)$$

$$S_{PEARSON-I} = \chi^2 \text{ where } \chi^2 = \frac{n(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)} \quad (51)$$

$$S_{PEARSON-II} = (\frac{\chi^2}{n+\chi^2})^{1/2} \quad (52)$$

$$S_{PEARSONIII} = (\frac{\rho}{n+\rho})^{1/2} \text{ where } \rho = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad (53)$$

$$S_{PEARSON\&HERON-I} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad (54)$$

$$S_{PEARSON\&HERON-II} = Cos(\frac{\pi\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}) \quad (55)$$

$$S_{SOKAL\&SNEATH-III} = \frac{a+d}{b+c} \quad (56)$$

$$S_{SOKAL\&SNEATH-V} = \frac{ad}{(a+b)(a+c)(b+d)(c+d)^{0.5}} \quad (57)$$

$$S_{COLE} = \frac{\sqrt{2}(ad-bc)}{\sqrt{(ad-bc)^2-(a+b)(a+c)(b+d)(c+d)}} \quad (58)$$

$$S_{STILES} = \log_{10}\frac{n(|ad-bc|-\frac{n}{2})^2}{(a+b)(a+c)(b+d)(c+d)} \quad (59)$$

$$S_{OCHIAI-II} = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad (60)$$

$$S_{YULEQ} = \frac{ad-bc}{ad+bc} \quad (61)$$

$$D_{YULEQ} = \frac{2bc}{ad+bc} \quad (62)$$

$$S_{YULEw} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}} \quad (63)$$

$$S_{KULCZYNSKI-I} = \frac{a}{b+c} \quad (64)$$

$$S_{TANIMOTO} = \frac{a}{(a+b)+(a+c)-a} \quad (65)$$

$$S_{DISPERSON} = \frac{ad-bc}{(a+b+c+d)^2} \quad (66)$$

$$S_{HAMANN} = \frac{(a+d)-(b+c)}{a+b+c+d} \quad (67)$$

$$S_{MICHAEL} = \frac{4(ad-bc)}{(a+d)^2+(b+c)^2} \quad (68)$$

$$S_{GOODMAN\&KRUSKAL} = \frac{\sigma-\sigma'}{2n-\sigma'} \text{ where}$$
$$\sigma = \max(a,b)+\max(c,d)+\max(a,c)+\max(b,d), \quad (69)$$
$$\sigma' = \max(a+c,b+d)+\max(a+b,c+d)$$

$$S_{ANDERBERG} = \frac{\sigma-\sigma'}{2n} \quad (70)$$

$$S_{BARONI-URBANI\&BUSER-I} = \frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c} \quad (71)$$

$$S_{BARONI-URBANI\&BUSER-II} = \frac{\sqrt{ad}+a-(b+c)}{\sqrt{ad}+a+b+c} \quad (72)$$

$$S_{PEIRCE} = \frac{ab+bc}{ab+2bc+cd} \quad (73)$$

$$S_{EYRAUD} = \frac{n^2(na-(a+b)(a+c))}{(a+b)(a+c)(b+d)(c+d)} \quad (74)$$

$$S_{TARANTULA} = \frac{\frac{a}{(a+b)}}{\frac{c}{(c+d)}} = \frac{a(c+d)}{c(a+b)} \quad (75)$$

$$S_{AMPLE} = \left|\frac{\frac{a}{(a+b)}}{\frac{c}{(c+d)}}\right| = \left|\frac{a(c+d)}{c(a+b)}\right| \quad (76)$$

The inclusion or exclusion of negative matches, *d* in the binary similarity measures have been an ongoing issue [9, 12, 15, 16, 17, 18, 26, 27]. The *Sokal & Michener*, the *Roger & Tanimoto*, the *Faith*, the *Ochiai II*, the *Cole*, the *Gower*, *Pearson I*, and the *Stiles* etc. are included in the *negative match inclusive* measures. The *Jaccard*, the *Tanimoto*, the *Dice & Sorenson*, the *Kulczynski I*, the *Ochiai I*, the *Mountford*, the *Sorgenfrei*, and the *Simpson* etc. are included in the *negative match exclusive* measures. Sokal et al. argued that the negative matches do not mean necessarily any similarity between two objects [27]. This is because an almost infinite number of attributes is possibly lacking in two objects.

In cases where the two binary states are not equally important, such as in the asymmetric type of binary data, the positive matches are usually more significant than the negative matches [1, 6, 10, 26]. Faith included the negative match but only gave the half credits while giving the full credits for the positive matches in eqn (10) [11]. In [4], different weights for positive and negative matches were studied. Weighted similarity measures such as weighted hamming distance or *azzoo* [4] are not covered in this paper though.

Historically, all the binary measures observed above have had a meaningful performance in their respective fields. The binary similarity coefficients proposed by Peirce, Yule, and Pearson in 1900s contributes to the evolution of the various correlation based binary similarity measures. The Jaccard coefficient proposed at 1901 is still widely used in the various fields such as ecology and biology. The discussion of inclusion or exclusion of *negative matches* was actively arisen by Sokal & Sneath in during 1960s and by Goodman & Kruskal in 1970s. In Figure 1, the measures are arranged in historical order.

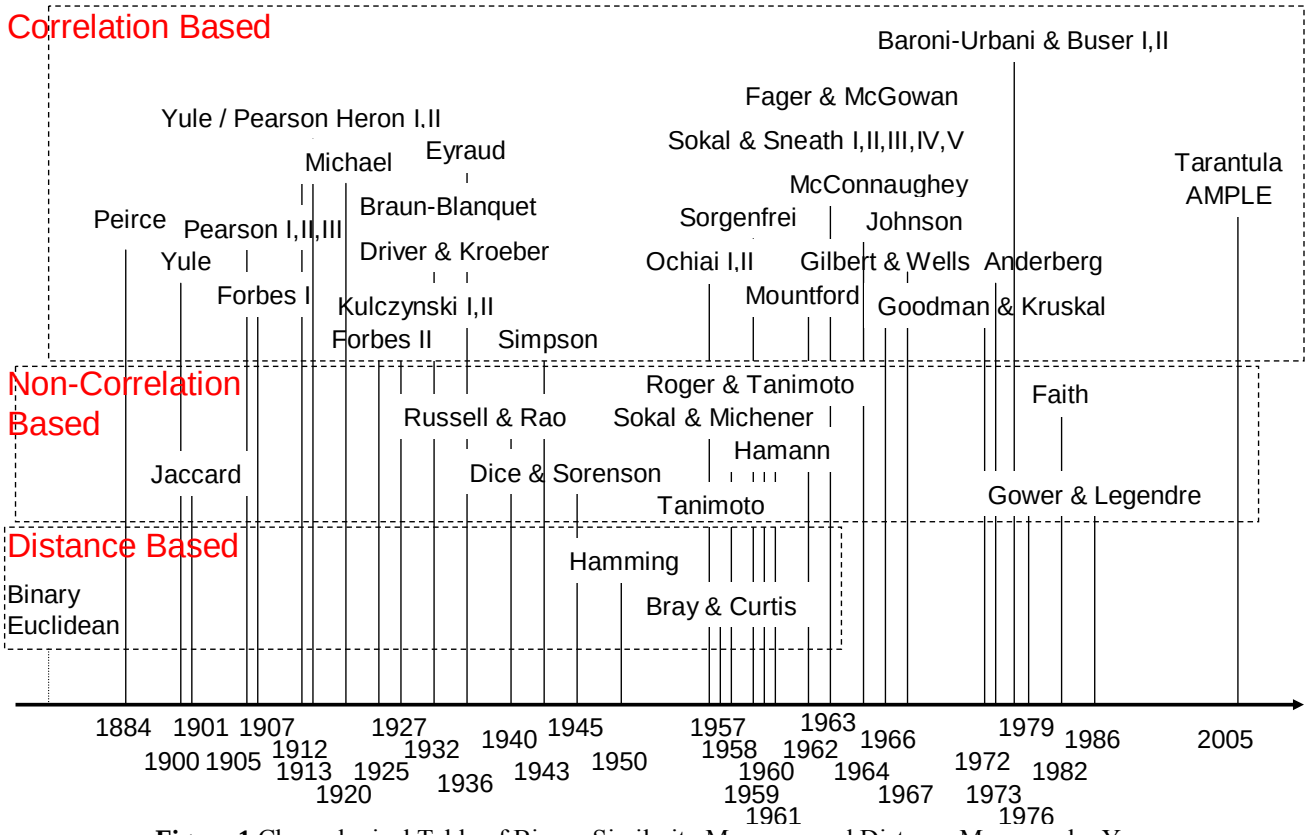

Figure 1 Chronological Table of Binary Similarity Measures and Distance Measures by Year

3. HIERARCHICAL CLUSTERING

Hierarchical clustering is conducted to estimate the similarity among the measures collected. Random binary data set are used as data set. The reference set consist of 30 binary instances, each of which has 100 binary features. When a test query is measured with the reference set data, 100 distance or similarity values are produced for each measure. The correlation coefficient values between two measures are used to build a *dendrogram*. The agglomerative single linkage with the average clustering method is used [9].

The *dendrogram* in Figure 2 is produced by averaging 30 independent trials. The vertical scale on the left side of dendrogram represents the binary similarity or dissimilarity measures examined. The horizontal scale represents the closeness of two clusters of binary similarity or dissimilarity measures, where $0 \le r \le 1$. The dendrogram provides intuitive semantic groupings of binary similarity measures and distance measures.

High correlations are found in the most of measures including negative matches. They are identified as Group 1 including the *Simple Matching*, the *Pearson*'s *phi-like* coefficients, and the *Yule Q*. The exceptional case is the Yule w, which has a square root of $ad - bc$ in the numerator. It is clustered in Group 5 showing different behavior. All of the *Hamming*-like binary distance measures are categorized in Group 1 while the *Lance & Williams* and the *Bray-Curtis* distance measures are clustered in Group 2 closely related with the *Hellinger* and the *Chord* distance measures. Most of negative match exclusive measures are clustered in Group 2 and 3. Additive form of negative match exclusive measures such as the *Jaccard*, the *Dice & Sorenson*, or the *Kulczynski I*, have high correlation with the *Cosine* based measures such as the *Ochiai I* or the *Sorgenfrei*. Interestingly, the *Faith* is categorized in Group 2 even though it is a variation of the *Sokal & Michener* of Group 1. The *Driver & Kroeber*, the *Forbes I*, and the *Fossum* have high correlation with *inner product* based measures such as the *Russell & Rao*. They are clustered in Group 3. The probabilistic similarity measures such as the *Goodman & Kruskal* and the *Anderberg* are identical as clustered in Group 6. The *Yule w*, the *Eyraud*, the *Fager & McGowan*, the *Stiles*, the *Tanimoto*, and the *Peirce* are different from others as they are clustered in Group 5, 7, 8, 9, 10, and 11 respectively. The *Chi-square* based measures such as the *Pearson I* and *Pearson II* are clustered separately forming Group 4. The *Tarantula* has high correlation with the *Sokal & Sneath III* and clustered in Group 1 while the *AMPLE* coefficient, the absolute value of the *Tarantula*, has high correlation with chi-square based measures and clustered in Group 4.

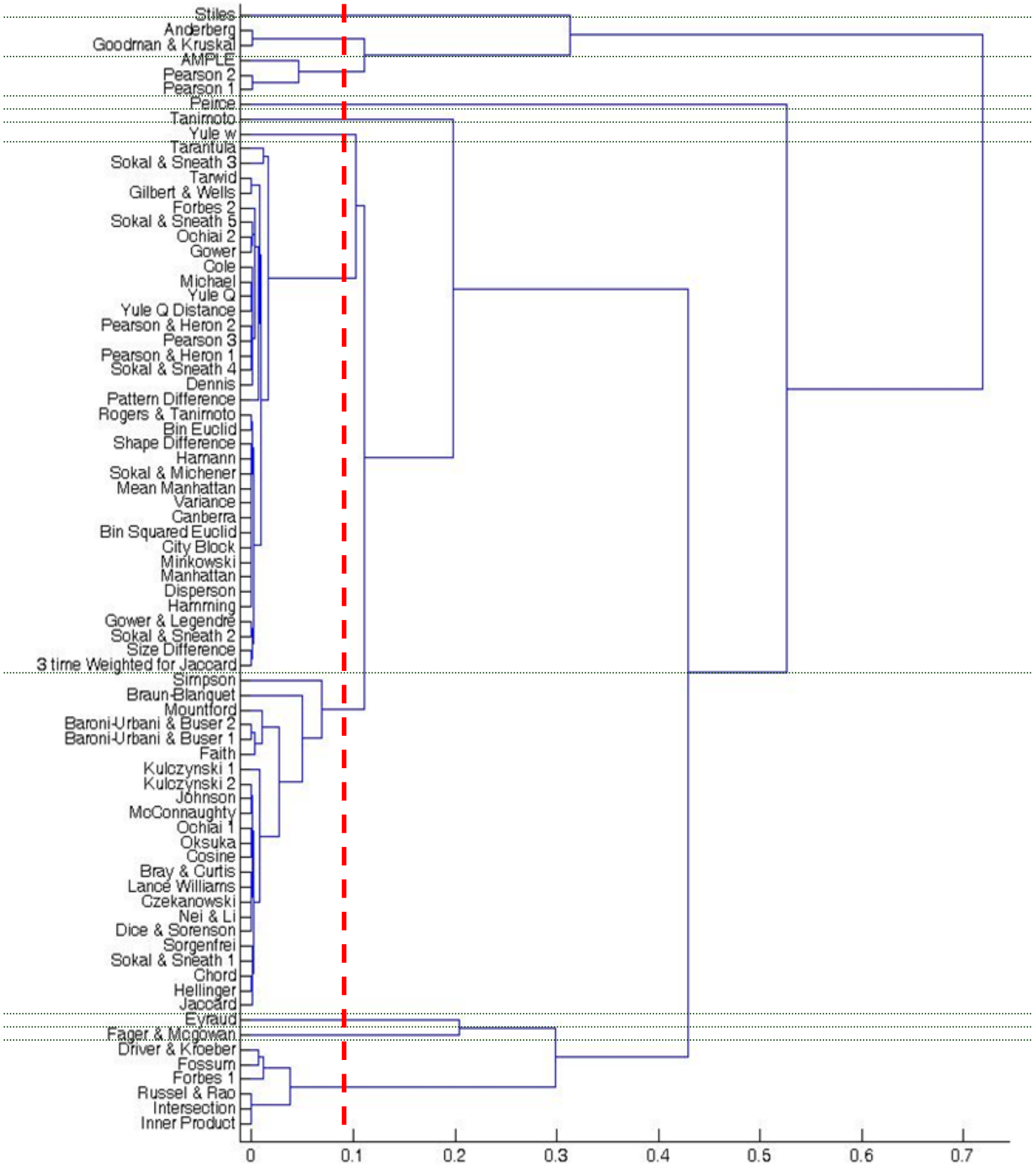

Figure 2 Hierarchical Clustering Result of Random Binary Data Set

4. CONCLUSIONS

Numerous binary similarity measures and distance measures have been used in various fields. Each of them is differently defined by its own synthetic properties. Some include negative matches and some do not. Some use simple count difference and some utilize complicated correlation. In this survey, we collected 76 binary similarity and distance measures used over the last century, classified them through hierarchical clustering, and observed close relationships among some of the measures. We expect that the relationship of each pair of measures should help researchers select more accurate measure for binary data analysis in various domains.

5. REFERENCES

[1] Baroni-Urbani, C., Buser, M.W., (1976), "Similarity of Binary Data", Systematic Zoology, Vol. 25, No. 3, pp. 251-259.

[2] Cha, S.-H., Srihari, S.N., (2000), "A fast nearest neighbor search algorithm by filtration", Pattern Recognition 35, P 515-525.

[3] Cha, S.-H., Tappert, C.C., (2003), "Optimizing Binary Feature Vector Similarity Measure using Genetic Algorithm", ICDAR, Edinburgh, Scotland.

[4] Cha, S.-H., Yoon S-, Tappert, C.C., (2006), "Enhancing Binary Feature Vector Similarity Measures", Journal of Pattern Recognition research I.

[5] Choi, S.-S, (2008), "Correlation Analysis of Binary Similarity Measures and Dissimilarity Measures", Doctorate dissertation, Pace University.

[6] Clifford, H., Stephenson, W., (1975), "An Introduction to Numerical Taxonomy", Academic Press, New York.

[7] Cormack, R.M., (1971), "A review of classification", Journal of the Royal Statistical Society, Series A, 134., pp. 321 - 353.

[8] Driver, H.E., Kroeber, A.L., (1932), "Quantitative Expression of Cultural Relationships", University of California Press.

[9] Dunn, G., Everitt, B.S., (1982), "An Introduction to Mathematical Taxonomy", Cambridge University Press.

[10] Faith, D.P, (1983), "Asymmetric binary similarity measures", Oecologia, Vol.57, No. 3, pp. 287-290.

[11] Faith, D.P., Minchin, P.R., Belbin, L., (1987), "Compositional dissimilarity as a robust measure of ecological distance", Journal of Plant Ecology, Volume 69, Numbers 1-3.

[12] Finely, J.P., (1884), "Tornado prediction," The American Meteorological Journal, 1, 85-8.

[13] Forbes, S.A., (1907), "On the local distribution of certain Illinois fishes. An essay in statistical ecology," Bulletin of the Illinois State Laboratory of Natural History.

[14] Forbes, S.A., (1925), "Method of determining and measuring the associative relations of species", Science 61, 524.

[15] Gilbert, G.K., (1884), "Finely's tornado predictions," The American Meteorological Journal, 1, 166-72.

[16] Goodman, L.A., Kruskal, W.H., (1954), "Measures of association for cross classifications", Journal of the American Statistical Association 49, 732-764.

[17] Goodman, L.A., Kruskal, W.H., (1959), "Measures of association for cross classifications II. Further discussion and references", Journal of the American Statistical Association 54, 123-163 (pp. 35-75).

[18] Goodman, L.A., Kruskal, W.H., (1963), "Measures of association for cross classifications III. Approximate sampling theory", Journal of the American Statistical Association 58, 310-364.

[19] Hubalek, Z., (1982), "Coefficients of Association and Similarity, Based on Binary (Presence-Absence) Data: An Evaluation", Biological Reviews, Vol.57-4,669-689.

[20] Jaccard, P., (1901), "Étude comparative de la distribuition florale dans une portion des Alpes et des Jura", Bull Soc Vandoise Sci Nat 37:547-579.

[21] Jackson, D.A., Somers, K.M., Harvey, H.H., (1989), "Similarity Coefficients: Measures of Co-Occurrence and Association or Simply Measures of Occurrence?", The American Nat1uralist, Vol. 133, No. 3, pp. 436-453.

[22] Kuhns, J.L., (1965), "The continuum of coefficients of association", Statistical Association Methods for Mechanized Documentation, (Edited by Stevens et al.) National Bureau of Standards, Washington, 33-39.

[23] Michael, E.L., (1920), "Marine ecology and the coefficient of association: a plea in behalf of quantitative biology", Ecology 8, 54-59.

[24] Michael H., (1976), "Binary coefficients: A theoretical and empirical study, Mathematical Geology, Volume 8, Number 2, April, 1976.

[25] Smith, J.R., Chang, S.-F., (1996), "Automated binary texture feature sets for image retrieval", International Conf. Accoust., Speech, Signal processing, Atlantic, GA.

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

[27] Sokal, R.R., Sneath P.H., (1963), "Principles of numeric taxonomy", San Francisco, W.H. Freeman.

[28] Tubbs, J.D., (1989), "A note on binary template matching", Pattern Recognition, 22(4):359-365.

[29] Willett, P., Barnard, J.M., Downs, G.M., (1998), "Chemical similarity searching" Chem Inf Comput Sci 38: 983-996.

[30] Willett, P., (2003), "Similarity-based approaches to virtual screening", Biochemical Society Transactions 31, 603–606.

[31] Zhang, B., Srihari, S.N., (2003), "Binary vector dissimilarities for handwriting identification", Proceedings of SPIE, Document Recognition and Retrieval X, p 15-166.