A Connection Between Formal Concept Analysis and Boolean Dissimilarities [1]

by Mel Janowitz, DIMACS [2]

[1] A preliminary version of a portion of this talk was given at Ecole Nationale Supérieure des Télécommunications de Bretagne on October 30, 2004

[2] Expanded version written 3/11/2005

Dissimilarity Coefficient (DC):
d: $P \times P \to \Re_0^+$ that satisfies
$d(a,b) = d(b,a)$ for all $a, b \in P$.
$d(a,a) = 0$ for all $a \in P$.

DC d is an ultrametric if
$d(a,b) \leq \max\{d(a,c), d(b,c)\}$
for all $a, b, c \in P$.

Idea: lower values of d correspond to pairs of P that are more similar (i.e., less dissimilar).

$\Sigma(P)$ = reflexive symmetric relations on P ordered by inclusion.

T-transform: Td: $\Re_0^+ \to \Sigma(P)$ defined by

$Td(h) = \{(a,b) : d(a,b) \leq h\}$ for all $h \in \Re_0^+$.

Easy to show $Td(h)$ is an equivalence relation for all $h \in \Re_0^+$ if and only if d is an ultrametric.

Note: $d \mapsto Td$ is a bijection.

Cluster algorithm may be viewed as a transformation $d \mapsto C(d)$ of a DC d into an ultrametric $C(d)$.

Generalization to a join semilattice L with 0:

$d\colon P \times P \to L$ satisfies

$d(a,b) = d(b,a)$ for all $a, b \in P$.

1. $d(a,a) = 0$ for all $a \in P$, or as alternates
2. $d(a,a) = \wedge\{d(a,b) : a \neq b,\ b \in P\}$,
3. apply the same formula as $d(a,b)$ with $b \neq a$.

DC d is an ultrametric if
$d(a,b) \leq d(a,c) \vee d(b,c)$ for all $a, b, c \in P$.

For L a Boolean algebra, d is called a Boolean dissimilarity.

Extra Note: If $d(x, x)$ is computed by any method that does not force it to be the 0 element of L, then we must replace $\Sigma(P)$ by the symmetric relations on P. It then follows that $Td(h)$ is transitive for all $h \in L$ if and only if d is an ultrametric. The point is that the ultrametric condition forces $d(a, b) \leq h$ whenever both $d(a, c)$ are $d(c, b)$ are $\leq h$,

Special type of Boolean DC:

Each element of P has k binary attributes.
So can represent
$a = (a_1, a_2, \ldots, a_k)$, $b = (b_1, b_2, \ldots, b_k)$, etc.
Want a DC d: $P \times P \to 2^k$, so

$$d(a, b) = (x_1, x_2, \ldots, x_k).$$

Since a DC is supposed to be a measure of how dissimilar a and b are, it is clear that if $a_i \neq b_i$, we want $x_i = 1$. There are now only two remaining cases to consider. $a_i = b_i = 0$ and $a_1 = b_i = 1$. The possible choices for x_i are:

$a_i = b_i = 0$, take $x_i = 0$

$a_i = b_i = 0$, take $x_i = 1$

$a_i = b_i = 1$, take $x_i = 0$

$a_i = b_i = 1$, take $x_i = 1$.

Of these four possibilities, we choose the following three.

$d_1(a, b) = (x_1, x_2, \ldots, x_k)$ where $x_i = 1$ if $a_i \neq b_i$ and 0 otherwise.

$d_2(a, b) = (y_1, y_2, \ldots, y_k)$ where $y_i = 0$ if $a_i = b_i = 1$ and 1 otherwise.

$d_3(a, b) = (z_1, z_2, \ldots, z_k)$ where $z_i = 0$ if $a_i = b_i = 0$ and 1 otherwise.

Theorem 1. *d_1, d_2, d_3 are each ultrametrics.*

Note: Need not consider d_3 as d_2 and d_3 are symmetric with respect to negation of attributes.

The three versions of $d(a, a)$ are all different.

Objects	$a1$	$a2$	$a3$	$a4$	$a5$
x	0	1	0	0	0
y	1	0	1	0	0
z	0	0	0	1	0

$$d_2(x, y) = d_2(x, z) = d_2(y, z) = 11111$$

Here are the three methods for finding $d_2(a, a)$.

	Method 1	Method 2	Method 3
$d_2(x, x)$	00000	11111	10111
$d_2(y, y)$	00000	11111	01011
$d_2(z, z)$	00000	11111	11101

Note: Easy to show that $d_1(a, a) = 0$ no matter which of the three methods of calculation is used.

If a denotes the number of shared presences between a pair of objects, c the number of shared absences, and b the number of attributes with only one presence, the classical coefficients corresponding to d_1 and d_2 are are computed as follows:

(d_1) Simple matching coefficient: $\dfrac{b}{a+b+c}$

(d_2) Russell and Rao coefficient: $\dfrac{b+c}{a+b+c}$.

Closing Example:

Consider the set P consisting of the first nine integers. We wish to classify P by considering various properties that these integers might enjoy. These properties are the *attributes* we shall consider. Here are some we might consider

o odd

s perfect square

p prime

c perfect cube

t multiple of three

This leads to the attributes presented in the Table below. The idea is that an entry of 1 indicates presence of the attribute, while 0 denotes absence. The data in the Table may be interpreted either as the input to a formal concept analysis problem or as the input to a cluster analysis problem. If the d_2 dissimilarity coefficient is calculated with $d_2(x, x)$ correctly calculated) the output using cluster analysis coincides with the lattice formed by the extents of the FCA approach.

object	o	s	p	c	t
1	1	1	0	1	0
2	0	0	1	0	0
3	1	0	1	0	1
4	0	1	0	0	0
5	1	0	1	0	0
6	0	0	0	0	1
7	1	0	1	0	0
8	0	0	0	1	0
9	1	1	0	0	1

Attributes for the first nine integers

Note: At level 11110, d_2 has cluster 369, the 1 entries for t, while d_1 has the bipartition 369, 124578.

Level 01110 has the common 1 entries for o and t: $\{3, 9\}$.

Level	d_1			d_2	Intents
11110	124578	369		369	$\{t\}$
11101	18	2345679		18	$\{c\}$
11011	14689	2357		2357	$\{p\}$
10111	149	235678		149	$\{s\}$
01111	13579	2468		13579	$\{o\}$
11100	18	2457	369		
11010	148	257	69		
11001	18	2357	469		
10110	14	2578	36		
10101	23567	49			
10011	149	2357	68		
01110	157	248	39	39	$\{o,t\}$
01101	246	3579			
01011	19	357	468	357	$\{o,p\}$
00111	19	268	357	19	$\{o,s\}$
11000	18	257	69		
10100	257	36			
10010	14	257			
10001	2357	49			
01100	24	39	57		
01010	48	57		3	$\{p,t\}$
01001	357	46			
00110	28	57		9	$\{o,s,t\}$
00101	26	357		1	$\{o,c\}$
00011	19	357	68		
10000	257				
01000	57				
00100	57				
00010	57				
00001	357				

d_2	1	2	3	4	5	6	7	8	9
1	00101	11111	01111	10111	01111	11111	01111	11101	00111
2	11111	11011	11011	11111	11011	11111	11011	11111	11111
3	01111	11011	01010	11111	01011	11110	01011	11111	01110
4	10111	11111	11111	10111	11111	11111	11111	11111	10111
5	01111	11011	01011	11111	01011	11111	01011	11111	01111
6	11111	11111	11110	11111	11111	11110	11111	11111	11110
7	01111	11011	01011	11111	01011	11111	01011	11111	01111
8	11101	11111	11111	11111	11111	11111	11111	11101	11111
9	00111	11111	01110	10111	01111	11110	01111	11111	00110

Table 1: Illustration of the d_2 coefficient for the nine integers

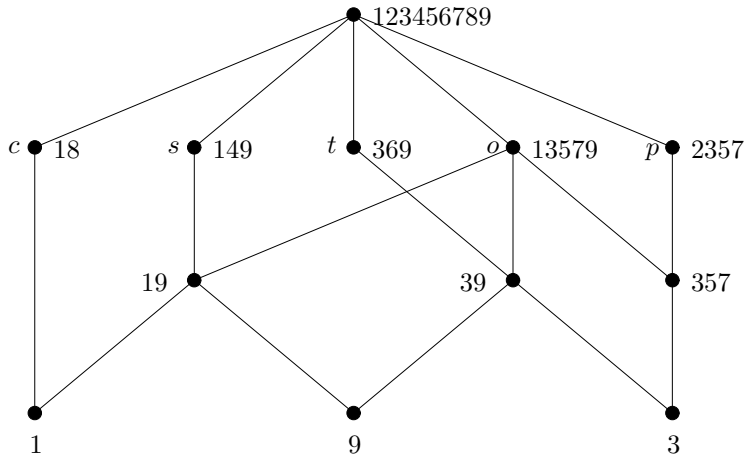

Figure 1: Nonempty extents of the nine integer example

Extra Page

d_1	1	2	3	4	5	6	7	8	9
1	00000	11110	01111	10010	01110	11011	01110	11100	00011
2	11110	00000	10001	01100	10000	00101	10000	00110	11101
3	01111	10001	00000	11101	00001	10100	00001	10111	01100
4	10010	01100	11101	00000	11100	01001	11100	01010	10001
5	01110	10000	00001	11100	00000	10101	00000	10110	01101
6	11011	00101	10100	01001	10101	00000	10101	00011	11000
7	01110	10000	00001	11100	00000	10101	00000	10110	01101
8	11100	00110	10111	01010	10110	00011	10110	00000	11011
9	00011	11101	01100	10001	01101	11000	01101	11011	00000

Table 2: Illustration of the d_1 coefficient for the nine integers

d_1	1	2	3	4	5	6	7	8	9
1	**00000**	11110	01111	10010	01110	11011	01110	**11100**	00011
2	11110	**00000**	10001	**01100**	**10000**	00101	**10000**	00110	11101
3	01111	10001	**00000**	11101	00001	**10100**	00001	10111	**01100**
4	10010	**01100**	11101	**00000**	**11100**	01001	**11100**	01010	10001
5	01110	**10000**	00001	**11100**	**00000**	10101	**00000**	10110	01101
6	11011	00101	**10100**	01001	10101	**00000**	10101	00011	**11000**
7	01110	**10000**	00001	**11100**	**00000**	10101	**00000**	10110	01101
8	**11100**	00110	10111	01010	10110	00011	10110	**00000**	11011
9	00011	11101	**01100**	10001	01101	**11000**	01101	11011	**00000**

Table 3: d_1 clusters at level 11100

Table 2 displays the clusters at level 11100. The entries of levels $\leq$ 11100 are displayed in boldface type. The clusters are 2457, 369, 18.

Extra Page

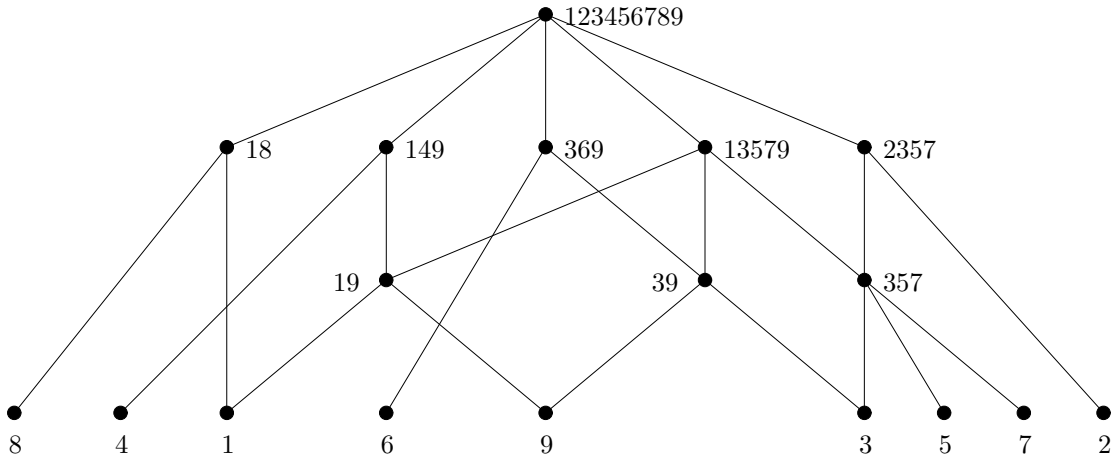

Figure 2: Clusters for the nine integer example with $d_2(x, x) = 0$.

The above figure illustrates the clusters formed when one takes $d_2(x, x) = 0$ for the nine integer example. The only change that being made is the addition of 6 new singleton clusters. This is in the spirit of cluster analysis rather than FCA, but the changes are trivial. The figure was drawn so as to be similar to Figure 1.

References

[DAV 90] B. A. Davey and H. A. Priestley, *Introduction to Lattices and Order*, Cambridge University Press, Cambridge, 1990.

[GAN 99] B. Ganter and R. Wille, *Formal Concept Analysis. Mathematical Foundations*, Springer-Verlag, Berlin, 1999.

[GOR 99] A. D. Gordon, *Classification, 2nd ed.*, Chapman & Hall, London, 1999.

[JAI 88] Nnil K. Jain and Richard C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs, 1988.

[JAN 78] M. F. Janowitz, *An order theoretic model for cluster analysis*, SIAM J. Applied Math. **34** (1978), 55-72.

[JAR 71] N. Jardine and R. Sibson, *Mathematical Taxonomy,* Wiley, New York, 1971. i

[MIR 96] B. Mirkin, *Mathematical Classification and Clustering*, Kluwer, Dodrecht, 1996.

[SNE 72] P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy*, W. H. Freeman, San Francisco, 1973.

[SZA 63] Gabor Szasz, *Introduction to Lattice Theory*, Third Ed., Academic Press, Budapest, 1963.