INTERNATIONAL CONFERENCE ON ENGINEERING DESIGN
ICED 97 TAMPERE, AUGUST 19-21, 1997

SIMILARITY AND EVALUATION OF TECHNICAL OBJECTS

Mehmet H. Göker
Technische Hochschule Darmstadt
Maschinenelemente und Konstruktionslehre
Magdalenenstr. 4
D-64289 Darmstadt, Germany
goker@muk.maschinenbau.th-darmstadt.de

Keywords *: similarity, evaluation, selection, retrieval, weighting factors*

1. Introduction

The similarity of a technical object to the ideal object described in the requirement definition document forms the basis for all evaluation techniques described in engineering design methodology [VDI2225-3, PaBe93]. This paper gives a brief description of similarity metrics that have been used in literature to measure the similarity of two objects [BaHaWe92] and, based on these, develops a new similarity metric that can be used both to evaluate the quality of technical objects and to select technical objects from a storage of available solutions. In contrast to the standard evaluation techniques used in engineering design methodology, the devised similarity metric takes the existence and absence of both attributes and properties in the ideal and the evaluated object into account [Gö96]. In addition, a method to enable the objective, effective and efficient determination of weighting factors used in this similarity metric is described.

2. Similarity Metrics and the Similarity of Technical Objects

The similarity of two objects is the degree to which their values for the attributes that define their common class match. Similarity can only be defined and determined in a given context and is neither transitive nor necessarily symmetric.

Similarity- and distance-metrics are used to measure and express similarity by numeric values. The two metric types are equivalent and can be transformed into each other. Given the maximum distance d_{max} and a distance function d(x) the similarity metric sim(x) can be calculated as [BaHeWe92] :

$$sim(x) = 1 - \frac{d(x)}{d_{\max}} \qquad \text{(Eq.1)}$$

x_i / y_i	1	0
1	$a = \sum_{i=1}^{n} (x_i y_i)$	$b = \sum_{i=1}^{n} (x_i \bar{y}_i)$
0	$c = \sum_{i=1}^{n} (\bar{x}_i y_i)$	$d = \sum_{i=1}^{n} (\bar{x}_i \bar{y}_i)$

Table 1:Contingency-table for property vectors with binary values

Whilst defining a similarity metric the matching and dissimilar properties of the compared objects have to be taken into account. In statistics, the contingency-table is used to express property correspondence and to define similarity metrics (Table 1) [Bo93, BaHeWe92].

The values shown in Table 1are calculated based on two property vectors $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ with binary values (i.e. $x_i, y_i \in \{0,1\}$).

In Table 1, a and d are the number of matching values, whereas b and c are the number of values that differ in

the two vectors. Based on this table, various similarity and distance metrics can be analysed and compared (Table 2). An extensive summary can be found in [BaHeWe92].

Metric and metric-type	Formula	Remarks
Hamming Distance (distance metric)	$d(x,y) = b + c$	Sum of differences
Simple Matching Coefficient (similarity metric)	$sim(x,y) = \frac{a+d}{a+b+c+d}$	Relative share of matches
S-Coefficient (similarity metric)	$sim(x,y) = \frac{a}{a+b+c}$	Relative share of positive matches

Table 2: Distance and similarity metrics for attributes with binary values

However, due to the fact that the validity of these metrics is limited to properties with values from nominal or ordinal scales (non-metric scales) they are not suitable for a general similarity analysis without modification.

On the other hand, for properties with values from interval or ratio scales (metric scales) similarity metrics based on the Euclidean distance are suggested [Bo93] (Table 3). In this table x and y are two property vectors $x = (x_1, x_2, ... , x_n)$ and $y = (y_1, y_2, ... , y_n)$ with values from metric scales.

Metric and metric type	Formula	Remarks
Euclidean Distance (Distance metric)	$d(x,y) = \sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2}$	w_i is the weighting factor of the respective property
City-Block Metric (Distance Metric)	$d(x,y) = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$	
Dominance Metric (Distance Metric)	$d(x,y) = \max_{i=1..n} \{ \lvert x_i - y_i \rvert \}$	
Minkowski r-Distance (Distance Metric)	$d(x,y) = \left(\sum_{i=1}^{n} \lvert x_i - y_i \rvert^r \right)^{\frac{1}{r}}$	$r \geq 1$; for r=1 this yields the City-Block Metric, for r=2 the Euclidean distance and for $r \to \infty$ the Dominance metric.

Table 3: Distance metrics for property vectors with values from metric scales

The restriction of value types is not appropriate for design tasks. Similarity in design is utilised to measure the suitability of an object under the given circumstances either in terms of an evaluation or for retrieval purposes. The requirements represent the properties of the ideal object against which the similarity of the current object is measured. The fact that the existence of non-required properties and absense of required properties is not taken into account is definitely a disadvantage of the metrics shown in Table 3. The contingency table takes this into account but does not distinguish between properties and attributes, i.e. if the property is not existent or required at all - or if it is only the value of the property that is different. Figure 1 shows an extended contingency table which takes this into account.

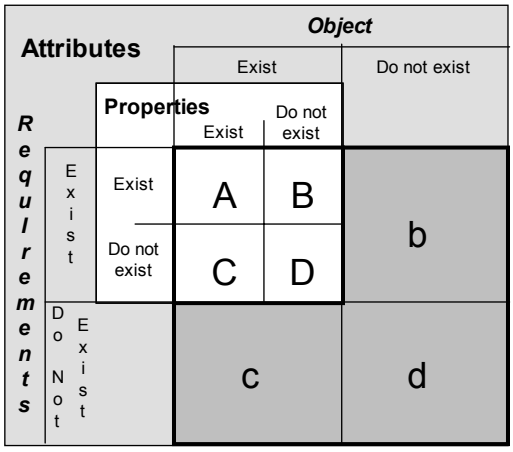

Figure 1:Extended contingency table

Similar to the contingency table for binary properties, the number of matching and dissimilar properties and attributes are calculated. A is the number of properties (attributes and corresponding values) that are requested in the requirements definition (description of the ideal object) and do exist in the current object. B and C are the properties that are requested but not present, or not requested but present in the current object. The difference between these two fields and b and c is, that in the latter two the attributes are not present (or requested) at all. D and d cover the same aspect and are per definition the attributes and properties that are irrelevant (Figure 2).

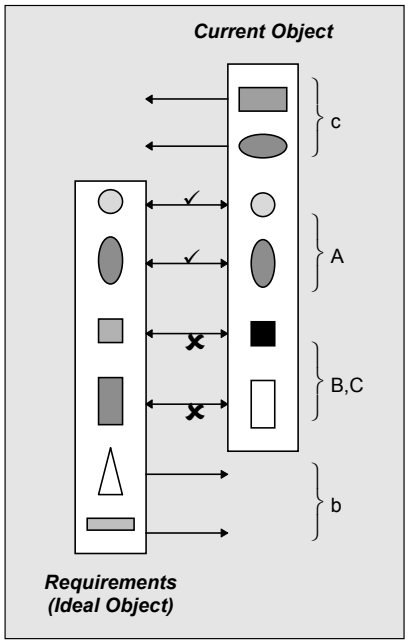

Figure 2:The values in the contingency table

A similarity metric that is suitable for design purposes has to take the values for A, B, C, c, and d into account. While the effect of A, B and C to the overall similarity of the two objects is of positive nature, b and c decrease the similarity. Obviously the effect of A is to be weighted higher than the effect of B and C. Based on these assumptions, a normalised, weighted similarity metric can be defined as follows (Eq.2):

$$sim(x,y) = \frac{w_A A + w_B B + w_C C - w_b b - w_c c}{A + B + C + b + c} \qquad \text{(Eq.2)}$$

The weighting factors ($w_A \ldots w_c$) have to be determined based on the relevant boundary conditions. It can be expected that for most cases $w_B=w_C$ and $w_b=w_c$.

However, as only the number of similar properties but not the distance to the required value for a each attribute is taken into account, this metric does not necessarily give a measure for the suitability of an object in the given situation, i.e. does not evaluate.

In order to obtain a more precise, evaluating metric, the definition has to be extended to incorporate normalised distance metrics and weighting factors for each attribute. If $d_i(x_i,y_i)$ is a normalised distance metric applicable to a given attribute and w_i is a normalised weighting factor for this attribute, A B and C can be defined as in Eq. 3 to 5.

$$A = \sum_{p=1}^{e} w_p(1 - d_p(x_p, y_p)) \qquad \text{(Eq. 3)}$$

$$B = \sum_{q=1}^{f} w_q(1 - d_q(x_q, y_q)) \qquad \text{(Eq. 4)}$$

$$C = \sum_{r=1}^{g} w_r(1 - d_r(x_r, y_r)) \qquad \text{(Eq. 5)}$$

Even if a complete match is required for any property in A and the distance should be zero, it make sense to facilitate small variations. As the number of attributes in each sum is different, the value for the upper limit of the sums varies. For b and c only the weighting factor for that attribute but not the distance metric for the attribute has to be taken into account. The evaluating similarity metric can now be defined as follows :

$$sim(x,y) = \frac{w_A \sum_{p=1}^{e} w_p(1 - d_p(x_p, y_p)) + w_B \sum_{q=1}^{f} w_q(1 - d_q(x_q, y_q)) + w_C \sum_{r=1}^{g} w_r(1 - d_r(x_r, y_r)) - w_b \sum_{s=1}^{h} w_s - w_c \sum_{t=1}^{k} w_t}{e + f + g + h + k} \quad \text{(Eq. 6)}$$

By selecting distance metrics which are suitable for the scale from which the attributes take their values, the restrictions for value types can be lifted.

3. Similarity and Evaluation

In engineering design methodology, the overall value V of an object is calculated as shown in (Eq. 7) where w_i is the weighting factor for each attribute and v_i a value function that maps the values of the attribute to a quality metric [VDI2225-3]. It is assumed that all attributes are present in both the ideal and the evaluated object.

$$V = \sum_{i=1}^{n} w_i v_i \quad \text{(Eq. 7)}$$

$$\sum_{i=1}^{k} w_i v_i = \sum_{i=1}^{k} w_i (1 - d_i(x_i, y_i)) \quad \text{(Eq. 8)}$$

$$\Leftrightarrow v_i = (1 - d_i(x_i, y_i)) \quad \text{(Eq. 9)}$$

Compared to this method, the similarity metric given above has the advantage of taking not just the properties that are present at every object but also the properties that are requested but absent or vice versa into account. As such, standard evaluation can be seen as a special case of the similarity metric shown in Eq. 6. Comparing Eq. 7 to the way A (properties that exist in both objects) is calculated (Eq. 3), we can see that the value function in standard evaluation is actually a similarity metric (Eq. 9).

4. The Calculation of Weighting Factors

Obviously one of the major problems in using the similarity metric described above and in evaluation in general is the determination of the weighting factors for each attribute taken into account, i.e. each evaluation criterion. Although the influence of weighting factors decreases rapidly with the number of attributes or criteria used, they can effect the outcome of any similarity calculation drastically. Finding an effective and efficient way to determine the weighting factors as objectively as possible is therefore of central importance.

Weighting factors can be determined by means of objective trees [PaBe93] or through weighting factor matrices. In objective trees the main goal or objective is divided into sub-goals successively. In each step, the weighting factor of the super-goal is divided between the sub-goals. The sum of all weighting factors at one level always equals one. This approach has the effect that the designer has to analyse the goals he wants to reach quite thoroughly. However, the weighting factors of sub-goals originating from different branches may be out of balance if compared to each other.

k / i	Cr_1	Cr_2	Cr_3	Cr_4	Cr_5
Cr_1	1	3	2	1	1
Cr_2	1/3	1	2/3	1/3	1/3
Cr_3	1/2	3/2	1	1/2	1/2
Cr_4	1	3	2	1	1
Cr_5	1	3	2	1	1

Figure 3: Relative importance of criteria

The weighting factor matrix bases the calculation of weighting factors on a comparison of criteria (see also [Br93], however the formulae given in that paper should be checked before being put to use). In this approach a matrix is created in which the relative importance of the criteria with respect to each other is noted (Figure 3).

Based on a hypothetical absolute importance M_i of a criterion Cr_i, the relative importance m_{ik} of criterion Cr_i with respect to criterion Cr_k is defined as in (Eq. 10).

In the sample shown in Figure 3, criterion 1 is three times as important as criterion 2, two times as important as criterion 3, and as important as criterion 4 and 5. Due to the definition of relative importance, the fields in the first column are also fixed (Eq. 11). As a criterion is always as important as itself the values on the diagonal are always one (Eq. 12).

$$m_{ik} = \frac{M_i}{M_k} \qquad \text{(Eq. 10)}$$

$$m_{ik} = \frac{1}{m_{ki}} \qquad \text{(Eq.11)}$$

$$m_{ii} = \frac{M_i}{M_i} = 1 \qquad \text{(Eq. 12)}$$

Based on the relative importance given in the first row the rest of the matrix can be determined based on (Eq. 16). To determine the weighting factors, the sum of each row of relative importance is calculated and normalised to one (Figure 4, Eq. 17).

$$m_{ik} = \frac{M_i}{M_k} = \frac{M_i}{M_n} \times \frac{M_n}{M_k} \qquad \text{(Eq. 13)}$$

$$\Leftrightarrow m_{ik} = m_{in} \times m_{nk} \qquad \text{(Eq. 14)}$$

$$\Leftrightarrow m_{ik} = \frac{m_{nk}}{m_{ni}} \qquad \text{(Eq. 15)}$$

$$\text{therefore} \qquad m_{ik} = \frac{m_{1k}}{m_{1i}} \qquad \text{(Eq. 16)}$$

$$w_i = \frac{\sum_k m_{ik}}{\sum_n \sum_k m_{nk}} \qquad \text{(Eq. 17)}$$

k / i	Cr_1	Cr_2	Cr_3	Cr_4	Cr_5	Σ	w_i
Cr_1	1	3	2	1	1	8	0,26
Cr_2	1/3	1	2/3	1/3	1/3	8/3	0,09
Cr_3	1/2	3/2	1	1/2	1/2	4	0,13
Cr_4	1	3	2	1	1	8	0,26
Cr_5	1	3	2	1	1	8	0,26

Figure 4:Calculation of weighting factors

If M_a is the hypothetical absolute importance of any criterion, the following formula for its weighting factor can be derived :

$$w_i = \frac{\sum_j m_{ij}}{\sum_n \sum_j m_{nj}} = \frac{\sum_j \frac{M_i}{M_j}}{\sum_n \sum_j \frac{M_n}{M_j}} \qquad \text{(Eq. 18)}$$

$$\Leftrightarrow w_i = \frac{M_i \sum_j \frac{1}{M_j}}{\sum_n M_n \sum_j \frac{1}{M_j}} \qquad \text{(Eq. 19)}$$

$$\Leftrightarrow w_i = \frac{M_i}{\sum_n M_n} \qquad \text{(Eq. 20)}$$

$$\Leftrightarrow w_i = \frac{\frac{M_i}{M_a}}{\frac{1}{M_a}\sum_n M_n} = \frac{\frac{M_i}{M_a}}{\sum_n \frac{M_n}{M_a}} \qquad \text{(Eq. 21)}$$

$$\Leftrightarrow w_i = \frac{m_{ia}}{\sum_n m_{na}} \qquad \text{(Eq. 22)}$$

Thus, weighting factors can be calculated by just filling out one column (or row) of relative importance in the weighting factor matrix. One criterion has to be set into relation with all other criteria to do this. Even if this will not be possible for all cases in reality, the missing values can be calculated using Eq. 15. The number of relative importance needed to calculate all weighting factors equals the number or criteria. Even though the method is based on a subjective evaluation of importance between criteria, it ensures the internal consistency between all weighting factors.

5. Results

Based on an extended contingency table, a similarity metric that takes both the properties that are present and absent in the requirements and the evaluated object into account was developed [Gö96]. The similarity metric extends the standard evaluation techniques used in

engineering design and is more suitable for the retrieval of objects from a solution base and the measurement of their applicability. In order to be able to determine the weighting factors used both in the newly developed similarity metric and in evaluation and retrieval in general, a method to determine these factors efficiently and effectively was shown. The method enables to calculate the weighting factors for all attributes or criteria based on the comparison of importance of one attribute with all others and ensures consistency among all weighting factors.

6. References

BaHaWe92 M. Bayer, B. Herbig, S. Weß, "Ähnlichkeit und Ähnlichkeitsmaße" in "Fallbasiertes Schließen, eine Übersicht, Band I-III, pp. 135-153, S. Weß, K.D. Althoff, F. Maurer, J. Paulokat, R. Praeger, O. Wendel (eds.), SEKI Working Paper SWP-92-08 (SFB), Fachbereich Informatik, Universität Kaiserslautern, 1992

Bo93 J. Bortz, "Statistik für Sozialwissenschaftler", Vierte Auflage, Springer Verlag Berlin, Heidelberg, New York, 1993

Br93 A. Breiing, "Neue Gesichtspunkte zur Gewichtung von Bewertungskriterien", Konstruktion 45 (1993), pp.171-175, Springer Verlag, 1993 (see also Konstruktion 45, p. 320)

Gö96 M.H. Göker, "Einbinden von Erfahrung in das Konstruktionsmethodische Vorgehen", Dissertation, Technische Hochschule Darmstadt, 1996, Fortschrittsberichte VDI, Reihe 1, Nr. 268, VDI-Verlag, Düsseldorf, 1996

PaBe93 G. Pahl, W. Beitz, "Konstruktionslehre", 3.Auflage, Springer Verlag, Berlin, Heidelberg, New York, 1993

VDI2225-3 Verein Deutscher Ingenieure, VDI-Richtlinie 2225, Blatt 3, Entwurf, "Konstruktionsmethodik, Technisch-Wirtschaftliches Konstruieren, Technisch - Wirtschaftliche Bewertung", Beuth Verlag, Berlin, 1990