Rough Inclusion Functions and Similarity Indices*

Anna Gomolińska[1] and Marcin Wolski[2]

[1] Białystok University, Computer Science Institute,
Sosnowa 64, 15-887 Białystok, Poland,
anna.gom@math.uwb.edu.pl
[2] Maria Curie-Skłodowska University, Dept. of Logic and Philosophy of Science,
Pl. Marii Curie-Skłodowskiej 4, 20-031 Lublin, Poland,
marcin.wolski@umcs.lublin.pl

Abstract. Rough inclusion functions are mappings considered in the rough set theory with which one can measure the degree of inclusion of a set in a set (and in particular, the degree of inclusion of an information granule in an information granule) in line with rough mereology. On the other hand, similarity indices are mappings in cluster analysis with which one can compare clusterings, and clustering methods with respect to similarity. In this article we investigate the relationships between rough inclusion functions and similarity indices.

Keywords: rough inclusion function, rough mereology, similarity index, cluster analysis, granular computing.

1 Introduction

In 1994, L. Polkowski and A. Skowron introduced the formal notion of a *rough inclusion*, making it a fundamental concept of *rough mereology* (see, e.g. [1–4]).[3] Rough inclusion may be interpreted as a ternary relation with which one can express the fact that a set of objects is to some degree included in the same or another set of objects. Rough mereology is a theory extending the Leśniewski mereology [6, 7] from a theory of being-a-part to a theory of being-a-part-to-degree. *Rough inclusion functions* (RIFs) are mappings with which one can measure the degree of inclusion of sets in sets and which comply with the axioms of rough inclusion. Since according to L. A. Zadeh's definition [8], an *information granule* is a clump of objects drawn together on the basis of indistinguishability, similarity or functionality, RIFs can be used in particular to measure the degree of inclusion of information granules in information granules. Hence, the concept of a RIF is fundamental not only for the rough set theory [5, 9] but also for the foundations and the development of granular computing [10, 11].

* Many thanks to the anonymous referees for interesting comments on the paper. All errors left are our sole responsibility.

[3] It is worthy to note that some ideas on rough inclusion were presented by Z. Pawlak in [5].

RIFs can be useful in the rough set theory and, more generally, in granular computing in many ways. First, they can be applied to compare sets (and information granules) with respect to inclusion. Secondly, they can be used to define rough membership functions [12] and various approximation operators as those in the Skowron – Stepaniuk approach (see, e.g. [13, 14] and other papers by the same authors), in the Ziarko variable-precision rough set model (see, e.g. [15, 16] and more recent papers), or in the decision-theoretic rough set model [17, 18]. RIFs can also be used to estimate the confidence (known as accuracy as well) and the coverage of decision rules and association rules (see, e.g. [19]). Another application of RIFs is graded semantics of formulas (see, e.g. [20]). An important application of RIFs is obviously their usage to compute the degree of similarity (nearness, closeness) between sets of objects and, in particular, between information granules. Some steps into this direction have already been made (see, e.g. [21, 4, 14]).

The *similarity indices* we are going to speak about are used in *cluster analysis* [22–24] to compare clusterings, and clustering methods with respect to how they are similar to (or dissimilar from) one another. Many of these similarity indices were originally designed to compare species with respect to their mutual similarity, given information about presence and/or absence of some features. A. N. Albatineh, M. Niewiadomska-Bugaj, and D. Michalko thoroughly examined 28 similarity indices known from the literature on classification and cluster analysis, from which 22 turned out to be different.[4] The results of their research on *correction for chance agreement* for similarity indices can be found, e.g. in [25]. In the present article we continue our earlier works [26, 27], where among other things, three similarity indices out of those 22 were derived from RIFs. Our actual goal is to show that all 22 similarity indices investigated in [25] can be obtained starting with the RIFs $\kappa^{\pounds}$, κ_1, and κ_2 only. This reveals one more connection between the rough set theory and cluster analysis.

The rest of the paper is organized as follows. In Sect. 2 we recall the notion of a rough inclusion function and the three particular RIFs mentioned above. In Sect. 3 we present the 22 similarity indices known from the literature and discussed in [25], and we characterize them one by one by means of the standard RIF $\kappa^{\pounds}$ or two other RIFs, viz. κ_1 and κ_2. The last section contains final remarks.

2 Rough Inclusion Functions

Rough inclusion functions (RIFs for short) are supposed to be mappings to measure the degree of inclusion of sets in sets and to comply with the axioms of rough inclusion. In detail, a rough inclusion function upon a non-empty set of objects U (in short, a RIF upon U or simply, a RIF) is a mapping $\kappa : \wp U \times \wp U \mapsto [0, 1]$, assigning to any pair of sets (X, Y) of elements of U, a number $\kappa(X, Y)$ from the unit interval $[0, 1]$ interpreted as the degree to which X is included in

[4] Some similarity indices were introduced more than once, under different names.

Y, and such that the conditions $\mathrm{rif}_1(\kappa)$ and $\mathrm{rif}_2^*(\kappa)$ are satisfied, where

$$\mathrm{rif}_1(\kappa) \overset{\mathrm{def}}{\Leftrightarrow} \forall X, Y \subseteq U.(\kappa(X,Y) = 1 \;\Leftrightarrow\; X \subseteq Y),$$
$$\mathrm{rif}_2^*(\kappa) \overset{\mathrm{def}}{\Leftrightarrow} \forall X, Y, Z \subseteq U.(\kappa(Y,Z) = 1 \;\Rightarrow\; \kappa(X,Y) \leq \kappa(X,Z)).$$

Condition $\mathrm{rif}_1(\kappa)$ expresses the fact that the set-theoretical inclusion of sets is the most perfect case of rough inclusion. When $\mathrm{rif}_1(\kappa)$ holds, condition $\mathrm{rif}_2^*(\kappa)$ will be equivalent with condition $\mathrm{rif}_2(\kappa)$ below:

$$\mathrm{rif}_2(\kappa) \overset{\mathrm{def}}{\Leftrightarrow} \forall X, Y, Z \subseteq U.(Y \subseteq Z \;\Rightarrow\; \kappa(X,Y) \leq \kappa(X,Z))$$

expressing monotonicity of κ in the second variable. In the literature, weaker versions of RIFs are considered as well, where $\mathrm{rif}_1(\kappa)$ is replaced by "a half of it". Then, $\mathrm{rif}_2^*(\kappa)$ and $\mathrm{rif}_2(\kappa)$ will define different classes of inclusion mappings (see, e.g. [28]).

In summary, any RIF κ upon U should satisfy $\mathrm{rif}_1(\kappa)$ and $\mathrm{rif}_2^*(\kappa)$ or, equivalently, $\mathrm{rif}_1(\kappa)$ and $\mathrm{rif}_2(\kappa)$. Among RIFs, various subclasses of mappings can be distinguished by adding new postulates to be satisfied. These can be, for instance,

$$\mathrm{rif}_3(\kappa) \overset{\mathrm{def}}{\Leftrightarrow} \forall \emptyset \neq X \subseteq U.\kappa(X,\emptyset) = 0,$$
$$\mathrm{rif}_4(\kappa) \overset{\mathrm{def}}{\Leftrightarrow} \forall X, Y \subseteq U.(\kappa(X,Y) = 0 \;\Rightarrow\; X \cap Y = \emptyset),$$
$$\mathrm{rif}_4^{-1}(\kappa) \overset{\mathrm{def}}{\Leftrightarrow} \forall \emptyset \neq X \subseteq U. \forall Y \subseteq U.(X \cap Y = \emptyset \;\Rightarrow\; \kappa(X,Y) = 0),$$
$$\mathrm{rif}_5(\kappa) \overset{\mathrm{def}}{\Leftrightarrow} \forall \emptyset \neq X \subseteq U. \forall Y \subseteq U.(\kappa(X,Y) = 0 \;\Leftrightarrow\; X \cap Y = \emptyset),$$
$$\mathrm{rif}_6(\kappa) \overset{\mathrm{def}}{\Leftrightarrow} \forall \emptyset \neq X \subseteq U. \forall Y \subseteq U. \kappa(X,Y) + \kappa(X,Y^c) = 1,$$
$$\mathrm{rif}_7(\kappa) \overset{\mathrm{def}}{\Leftrightarrow} \forall X, Y, Z \subseteq U.(Z \subseteq Y \subseteq X \;\Rightarrow\; \kappa(X,Z) \leq \kappa(Y,Z)),$$

where Y^c denotes the set-theoretical complement of Y.[5] Obviously, $\mathrm{rif}_5(\kappa)$ if and only if $\mathrm{rif}_4(\kappa)$ and $\mathrm{rif}_4^{-1}(\kappa)$. Apart from that

$$\mathrm{rif}_4^{-1}(\kappa) \;\Rightarrow\; \mathrm{rif}_3(\kappa),$$
$$\mathrm{rif}_1(\kappa) \;\&\; \mathrm{rif}_6(\kappa) \;\Rightarrow\; \mathrm{rif}_5(\kappa). \tag{1}$$

The *standard* RIF, denoted by $\kappa^{\pounds}$ here, is the most famous and frequently used by the rough set community. The idea underlying this notion is closely related to the conditional probability. In logic, J. Łukasiewicz was the first who employed this idea when calculating the probability of truth associated with implicative formulas [31, 32]. Let us recall that $\kappa^{\pounds}$ is only defined for a finite U by putting

$$\kappa^{\pounds}(X,Y) \overset{\mathrm{def}}{=} \begin{cases} \frac{\#(X \cap Y)}{\#X} & \text{if } X \neq \emptyset, \\ 1 & \text{otherwise,} \end{cases} \tag{2}$$

[5] The last condition was mentioned in [29, 30]. There, rough inclusion is understood in a different way than in our paper.

where X, Y are any subsets of U and $\#X$ denotes the number of elements of X. In words, the standard RIF measures the fraction of the elements having the property described by the second argument (Y) among the elements with the property described by the first argument (X). Apart from being a true RIF, $\kappa^{\pounds}$ has a number of interesting properties recalled, e.g. in [27]. For instance, it satisfies $\mathrm{rif}_i(\kappa)$ $(i = 3, \ldots, 7)$ and $\mathrm{rif}_4^{-1}(\kappa)$.

Examples of other RIFs are mappings κ_1 and κ_2 such that for any $X, Y \subseteq U$,

$$\kappa_1(X,Y) \stackrel{\mathrm{def}}{=} \begin{cases} \frac{\#Y}{\#(X \cup Y)} & \text{if } X \cup Y \neq \emptyset, \\ 1 & \text{otherwise,} \end{cases}$$

$$\kappa_2(X,Y) \stackrel{\mathrm{def}}{=} \frac{\#(X^c \cup Y)}{\#U}. \tag{3}$$

Also in this case, U has to be finite. While κ_1 was introduced in [26], κ_2 had already been mentioned in [33]. The both RIFs were investigated in detail in [27]. The RIFs $\kappa^{\pounds}$, κ_1, and κ_2 are different from one another. Below we recall a few other properties of these mappings.

Proposition 1. *For any $X, Y \subseteq U$, we have:*

(i) $X \neq \emptyset \Rightarrow (\kappa_1(X,Y) = 0 \Leftrightarrow Y = \emptyset)$,
(ii) $\kappa_2(X,Y) = 0 \Leftrightarrow X = U \,\&\, Y = \emptyset$,
(iii) $\mathrm{rif}_4(\kappa_1)$ & $\mathrm{rif}_4(\kappa_2)$,
(iv) $\kappa^{\pounds}(X,Y) \leq \kappa_1(X,Y) \leq \kappa_2(X,Y)$,
(v) $\kappa_1(X,Y) = \kappa^{\pounds}(X \cup Y, Y)$ & $\kappa^{\pounds}(X,Y) = \kappa_1(X, X \cap Y)$,
(vi) $\kappa_2(X,Y) = \kappa^{\pounds}(U, X^c \cup Y)$.

Let us also note that due to (i), $\mathrm{rif}_3(\kappa_1)$ holds. The same cannot be however said about κ_2 (compare (ii)).

3 Similarity Indices in Terms of RIFs

In this section we reformulate the similarity indices studied in [25] in terms of the RIFs $\kappa^{\pounds}$, κ_1, or κ_2. The proofs that the indices can really be expressed in this way will be given in the full version of this paper.

Consider a set U_0 of $m > 0$ data points to be grouped by some clustering methods A_1 and A_2. Let U (our universe) be the set of all unordered pairs of data points $\{x, y\} \subseteq U_0$ to be compared in order to obtain clusterings, i.e. partitions of U_0 generated by A_1 and by A_2, and denoted by C_1 and C_2 here. Thus, $\#U = M = \binom{m}{2} = m(m-1)/2$. The similarity between the clusterings C_1 and C_2 (and the clustering methods A_1 and A_2) is usually assessed on the basis of the number of pairs of data points that are put into the same cluster or are put into different clusters by each of the grouping methods considered. For $i = 1, 2$, let us define

$$X_i = \{\{x, y\} \in U \mid x, y \text{ are clustered by } A_i\}. \tag{4}$$

Additionally, let

$$\begin{aligned} a &= \#(X_1 \cap X_2), \\ b &= \#(X_1 \cap X_2^c), \\ c &= \#(X_1^c \cap X_2), \\ d &= \#(X_1^c \cap X_2^c). \end{aligned} \tag{5}$$

In words, a is the number of pairs of data points $\{x, y\}$ such that x and y are placed in the same cluster according to both A_1 and A_2; b (respectively, c) is the number of pairs of data points $\{x, y\}$ such that x and y are placed in the same cluster by A_1 (resp., A_2), but they are placed in different clusters by A_2 (resp., A_1); finally, d is the number of pairs of data points $\{x, y\}$ such that x and y are placed in different clusters according to both A_1 and A_2. We also have $\#X_1 = a+b$, $\#X_2 = a+c$, $\#X_1^c = c+d$, $\#X_2^c = b+d$, and $\#U = a+b+c+d = M$. For simplicity assume that $a, b, c, d > 0$. Then, we will have that

$$\begin{aligned} \kappa^{\pounds}(X_1, X_2) &= \frac{a}{a+b}, \\ \kappa_1(X_1, X_2) &= \frac{a+c}{a+b+c}, \\ \kappa_2(X_1, X_2) &= \frac{a+c+d}{M}. \end{aligned} \tag{6}$$

In what follows we will present similarity indices one by one and their new formulation in terms of $\kappa^{\pounds}$, κ_1, or κ_2.

Wallace (1983). The similarity indices W_1, W_2 with range $[0, 1]$ were introduced by D. L. Wallace:

$$\begin{aligned} W_1(C_1, C_2) &\stackrel{\text{def}}{=} \frac{a}{a+b}, \\ W_2(C_1, C_2) &\stackrel{\text{def}}{=} \frac{a}{a+c}. \end{aligned} \tag{7}$$

It is easy to see that

$$\begin{aligned} W_1(C_1, C_2) &= \kappa^{\pounds}(X_1, X_2), \\ W_2(C_1, C_2) &= \kappa^{\pounds}(X_2, X_1). \end{aligned} \tag{8}$$

Kulczyński (1927). The similarity index K with range $[0, 1]$ was proposed by S. Kulczyński in 1927:

$$K(C_1, C_2) \stackrel{\text{def}}{=} \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right) \tag{9}$$

K can be rewritten to the following form:

$$K(C_1, C_2) = \frac{1}{2}\left(\kappa^{\pounds}(X_1, X_2) + \kappa^{\pounds}(X_2, X_1)\right) \tag{10}$$

In words, $K(C_1, C_2)$ is the arithmetical mean of $\kappa^{\pounds}(X_1, X_2)$ and $\kappa^{\pounds}(X_2, X_1)$.

McConnaughey (1964). The similarity index MC with range $[-1, 1]$ goes back to B. H. McConnaughey:

$$MC(C_1, C_2) \stackrel{\text{def}}{=} \frac{a^2 - bc}{(a+b)(a+c)} \tag{11}$$

This index can be expressed by the following equation:

$$MC(C_1, C_2) = \kappa^{\mathcal{L}}(X_1, X_2) + \kappa^{\mathcal{L}}(X_2, X_1) - 1 \tag{12}$$

Peirce (1884). The similarity index PE with range $[-1, 1]$ is attributed to C. S. Peirce:

$$PE(C_1, C_2) \stackrel{\text{def}}{=} \frac{ad - bc}{(a+c)(b+d)} \tag{13}$$

The index PE can be characterized as follows:

$$PE(C_1, C_2) = \frac{1}{2}\left(\kappa^{\mathcal{L}}(X_2, X_1) + \kappa^{\mathcal{L}}(X_2^c, X_1^c) - \kappa^{\mathcal{L}}(X_2, X_1^c) - \kappa^{\mathcal{L}}(X_2^c, X_1)\right) \tag{14}$$

The Gamma index. The similarity index Γ with range $[-1, 1]$ is given by

$$\Gamma(C_1, C_2) \stackrel{\text{def}}{=} \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}. \tag{15}$$

In this case, the following characterization can be obtained:

$$\begin{aligned}\Gamma(C_1, C_2) = &\sqrt{\frac{1}{2}\left(\kappa^{\mathcal{L}}(X_2, X_1) + \kappa^{\mathcal{L}}(X_2^c, X_1^c) - \kappa^{\mathcal{L}}(X_2, X_1^c) - \kappa^{\mathcal{L}}(X_2^c, X_1)\right)} \\ &\cdot \sqrt{\kappa^{\mathcal{L}}(X_1, X_2) - \kappa^{\mathcal{L}}(X_1^c, X_2)}\end{aligned} \tag{16}$$

Ochiai (1957), Fowlkes and Mallows (1983). The similarity index OFM ranges over $[0, 1]$. It was introduced by A. Ochiai in 1957 and again by E. B. Fowlkes and C. L. Mallows in 1983:

$$OFM(C_1, C_2) \stackrel{\text{def}}{=} \frac{a}{\sqrt{(a+b)(a+c)}} \tag{17}$$

After rewriting we get

$$OFM(C_1, C_2) = \sqrt{\kappa^{\mathcal{L}}(X_1, X_2)\kappa^{\mathcal{L}}(X_2, X_1)}. \tag{18}$$

That is, $OFM(C_1, C_2)$ is the geometrical mean of $\kappa^{\mathcal{L}}(X_1, X_2)$ and $\kappa^{\mathcal{L}}(X_2, X_1)$.

The Pearson index. The similarity index P named after C. Pearson ranges over $[-1, 1]$. It is given by

$$P(C_1, C_2) \stackrel{\text{def}}{=} \frac{ad - bc}{(a+b)(a+c)(b+d)(c+d)}. \tag{19}$$

The index P can be expressed in the following ways:

$$\begin{aligned} P(C_1, C_2) &= \begin{vmatrix} a & b \\ c & d \end{vmatrix}^{-1} \cdot \Gamma^2(C_1, C_2) \\ &= (\kappa^{\pounds}(X_1, X_2) - \kappa^{\pounds}(X_1^c, X_2))\kappa^{\pounds}(X_2, \{u\})\kappa^{\pounds}(X_2^c, \{u'\}) \end{aligned} \tag{20}$$

for arbitrary $u \in X_2$ and $u' \notin X_2$.

Sokal and Sneath (1963). The similarity indices SS_1, SS_2, SS_3 with range $[0, 1]$ were introduced by R. R. Sokal and P. H. Sneath in 1963. The third index is also attributed to A. Ochiai (1957):

$$\begin{aligned} SS_1(C_1, C_2) &\stackrel{\text{def}}{=} \frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right), \\ SS_2(C_1, C_2) &\stackrel{\text{def}}{=} \frac{a}{a + 2(b+c)}, \\ SS_3(C_1, C_2) &\stackrel{\text{def}}{=} \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}. \end{aligned} \tag{21}$$

One can prove the following:

$$\begin{aligned} SS_1(C_1, C_2) &= \frac{1}{4}\left(\kappa^{\pounds}(X_1, X_2) + \kappa^{\pounds}(X_2, X_1) + \kappa^{\pounds}(X_1^c, X_2^c) + \kappa^{\pounds}(X_2^c, X_1^c)\right), \\ SS_2(C_1, C_2) &= \frac{\kappa_1(X_1, X_2) + \kappa_1(X_2, X_1) - 1}{3 - (\kappa_1(X_1, X_2) + \kappa_1(X_2, X_1))}, \\ SS_3(C_1, C_2) &= \sqrt{\kappa^{\pounds}(X_1, X_2)\kappa^{\pounds}(X_2, X_1)\kappa^{\pounds}(X_1^c, X_2^c)\kappa^{\pounds}(X_2^c, X_1^c)}. \end{aligned} \tag{22}$$

Thus, $SS_1(C_1, C_2)$ (resp., $SS_3(C_1, C_2)$) is the arithmetical (geometrical) mean of $\kappa^{\pounds}(X_1, X_2)$, $\kappa^{\pounds}(X_2, X_1)$, $\kappa^{\pounds}(X_1^c, X_2^c)$, and $\kappa^{\pounds}(X_2^c, X_1^c)$.

Jaccard (1908). The similarity index J with range $[0, 1]$ goes back to P. Jaccard:

$$J(C_1, C_2) \stackrel{\text{def}}{=} \frac{a}{a+b+c} \tag{23}$$

It can be shown that

$$J(C_1, C_2) = \kappa_1(X_1, X_2) + \kappa_1(X_2, X_1) - 1. \tag{24}$$

Sokal and Michener (1958), Rand (1971). The similarity index R with range $[0,1]$ was introduced by R. R. Sokal and C. D. Michener, and later independently by W. Rand:

$$R(C_1, C_2) \stackrel{\text{def}}{=} \frac{a+d}{M} \tag{25}$$

The index R can be rewritten to

$$R(C_1, C_2) = \kappa_2(X_1, X_2) + \kappa_2(X_2, X_1) - 1. \tag{26}$$

Hamann (1961), Hubert (1977). The similarity index H, ranging over $[-1,1]$, was proposed by U. Hamann and independently by L. J. Hubert:

$$H(C_1, C_2) \stackrel{\text{def}}{=} \frac{(a+d)-(b+c)}{M} \tag{27}$$

By certain transformations we obtain

$$H(C_1, C_2) = 2(\kappa_2(X_1, X_2) + \kappa_2(X_2, X_1)) - 3. \tag{28}$$

Czekanowski (1932), Dice (1945), Gower and Legendre (1986). The similarity index CZ ranges over $[0,1]$. It was proposed by J. Czekanowski in 1932, L. R. Dice in 1945, and by J. C. Gower and P. Legendre in 1986:

$$CZ(C_1, C_2) \stackrel{\text{def}}{=} \frac{2a}{2a+b+c} \tag{29}$$

On can prove the following:

$$CZ(C_1, C_2) = \frac{2(\kappa_1(X_1, X_2) + \kappa_1(X_2, X_1) - 1)}{\kappa_1(X_1, X_2) + \kappa_1(X_2, X_1)} \tag{30}$$

Russel and Rao (1940). The similarity index RR ranges over $[0,1]$ and is attributed to P. F. Russel and T. R. Rao:

$$RR(C_1, C_2) \stackrel{\text{def}}{=} \frac{a}{M} \tag{31}$$

In this case we obtain that

$$RR(C_1, C_2) = \kappa^{\mathcal{L}}(U, X_1 \cap X_2) = \kappa_2(U, X_1 \cap X_2). \tag{32}$$

Fager and McGowan (1963). The similarity index FMG with range $[-1/2, 1)$ goes back to E. W. Fager and J. A. McGowan :

$$FMG(C_1, C_2) \stackrel{\text{def}}{=} \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{a+b}} \tag{33}$$

The above formula can be expressed in the following way:

$$FMG(C_1, C_2) = \sqrt{\kappa^{\mathcal{L}}(X_1, X_2)\kappa^{\mathcal{L}}(X_2, X_1)} - \frac{1}{2}\sqrt{\kappa^{\mathcal{L}}(X_1, \{u\})} \tag{34}$$

for an arbitrary $u \in X_1$.

Sokal and Sneath (1963), Gower and Legendre (1986). The similarity index GL with range $[0,1]$ was introduced by R. R. Sokal and P. H. Sneath in 1963, and again by J. C. Gower and P. Legendre in 1986:

$$GL(C_1,C_2) \stackrel{\text{def}}{=} \frac{a+d}{a+\frac{1}{2}(b+c)+d} \tag{35}$$

A characterization of GL in terms of κ_2 is the following:

$$GL(C_1,C_2) = \frac{2(\kappa_2(X_1,X_2)+\kappa_2(X_2,X_1)-1)}{\kappa_2(X_1,X_2)+\kappa_2(X_2,X_1)} \tag{36}$$

Rogers and Tanimoto (1960). The similarity index RT with range $[0,1]$ is attributed to D. J. Rogers and T. T. Tanimoto:

$$RT(C_1,C_2) \stackrel{\text{def}}{=} \frac{a+d}{a+2(b+c)+d} \tag{37}$$

This index can be rewritten to the following form:

$$RT(C_1,C_2) = \frac{\kappa_2(X_1,X_2)+\kappa_2(X_2,X_1)-1}{3-(\kappa_2(X_1,X_2)+\kappa_2(X_2,X_1))} \tag{38}$$

Yule (1927), Goodman and Kruskal (1954). The similarity index GK ranges over $[-1,1]$. It was proposed by G. U. Yule in 1927, and again by L. A. Goodman and W. H. Kruskal in 1954:

$$GK(C_1,C_2) \stackrel{\text{def}}{=} \frac{ad-bc}{ad+bc} \tag{39}$$

This index can be expressed in terms of the standard RIF as follows:

$$GK(C_1,C_2) = \frac{\kappa^{\mathcal{L}}(X_2,X_1)\kappa^{\mathcal{L}}(X_2^c,X_1^c)-\kappa^{\mathcal{L}}(X_2,X_1^c)\kappa^{\mathcal{L}}(X_2^c,X_1)}{\kappa^{\mathcal{L}}(X_2,X_1)\kappa^{\mathcal{L}}(X_2^c,X_1^c)+\kappa^{\mathcal{L}}(X_2,X_1^c)\kappa^{\mathcal{L}}(X_2^c,X_1)} \tag{40}$$

Baulieu (1989). The similarity indices B_1 and B_2 range over $[0,1]$ and $[-1/4,1/4]$, respectively. They were introduced by F. B. Baulieu in 1989:

$$\begin{aligned} B_1(C_1,C_2) &\stackrel{\text{def}}{=} \frac{M^2-M(b+c)+(b-c)^2}{M^2}, \\ B_2(C_1,C_2) &\stackrel{\text{def}}{=} \frac{ad-bc}{M^2}. \end{aligned} \tag{41}$$

As in all previous cases, a RIF (precisely, κ_2 here) underlies the definitions of these similarity indices, viz.,

$$\begin{aligned} B_1(C_1,C_2) &= \kappa_2(X_1,X_2)+\kappa_2(X_2,X_1)-1+(\kappa_2(U,X_1)-\kappa_2(U,X_2))^2, \\ B_2(C_1,C_2) &= (1-\kappa_2(X_1,X_2^c))\kappa_2(U,X_1^c)-(1-\kappa_2(X_1^c,X_2^c))\kappa_2(U,X_1). \end{aligned} \tag{42}$$

4 Final Remarks

The main goal realized in this paper was to show that a pretty vast number of various similarity indices known from the literature can be formulated in terms of some rough inclusion functions. Rough inclusion functions (RIFs) are mappings, inspired by the notion of a rough inclusion introduced by L. Polkowski and A. Skowron as a basic concept of rough mereology, by means of which one can measure the degree of inclusion of a set of objects in a set of objects. Since information granules can be viewed as particular sets of objects, RIFs are important not only for the rough set theory but also for granular computing.

Starting with the standard RIF $\kappa^{\pounds}$ and two other RIFs of a similar origin, denoted by κ_1 and κ_2, we have obtained all 22 similarity indices discussed in [25]. In the paper just mentioned it is proved that the indices K and MC are equivalent after some correction known as the correction for agreement due to chance, and the same holds for R, H, and CZ. We have not referred to this question because we are interested in other aspects concerning similarity indices. For example, we think about a usage of similarity indices in granular computing to calculate the degree of similarity between compound information granules such as indistinguishability relations and tolerance relations on a set of elementary objects considered. Let us note that similarity indices can also be used in granular computing in a more general setting, viz. to compute the degree of similarity between arbitrary sets of objects.

In the full version of this article we will give an illustrating example and proofs of the formulas characterizing the similarity indices considered. In the future research we will generalize our results, viz. we will propose general schemata for generation of similarity indices from an arbitrary RIF. Another question, also suggested by the referee, is the discovery of relationships among RIFs and quality measures for clusters.

References

1. Polkowski, L.: Reasoning by Parts: An Outline of Rough Mereology, Warszawa (2011)
2. Polkowski, L., Skowron, A.: Rough mereology. Lecture Notes in Artificial Intelligence **869** (1994) 85–94
3. Polkowski, L., Skowron, A.: Rough mereology: A new paradigm for approximate reasoning. Int. J. Approximated Reasoning **15**(4) (1996) 333–365
4. Polkowski, L., Skowron, A.: Rough mereological calculi of granules: A rough set approach to computation. Computational Intelligence **17**(3) (2001) 472–492
5. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning About Data. Kluwer, Dordrecht (1991)
6. Leśniewski, S.: Foundations of the General Set Theory 1 (in Polish). Volume 2 of Works of the Polish Scientific Circle., Moscow (1916) Also in [7], pages 128–173.
7. Surma, S.J., Srzednicki, J.T., Barnett, J.D., eds.: Stanisław Leśniewski Collected Works. Kluwer/Polish Scientific Publ., Dordrecht/Warsaw (1992)
8. Zadeh, L.A.: Outline of a new approach to the analysis of complex system and decision processes. IEEE Trans. on Systems, Man, and Cybernetics **3** (1973) 28–44

9. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences **177**(1) (2007) 3–27
10. Pedrycz, W., Skowron, A., Kreinovich, V., eds.: Handbook of Granular Computing. John Wiley & Sons, Chichester (2008)
11. Stepaniuk, J.: Rough-Granular Computing in Knowledge Discovery and Data Mining. Springer-V., Berlin Heidelberg (2008)
12. Pawlak, Z., Skowron, A.: Rough membership functions. In Fedrizzi, M., Kacprzyk, J., Yager, R.R., eds.: Fuzzy Logic for the Management of Uncertainty. John Wiley & Sons, New York (1994) 251–271
13. Skowron, A., Stepaniuk, J.: Tolerance approximation spaces. Fundamenta Informaticae **27**(2–3) (1996) 245–253
14. Stepaniuk, J.: Knowledge discovery by application of rough set models. In: [34]. (2001) 137–233
15. Ziarko, W.: Variable precision rough set model. J. Computer and System Sciences **46**(1) (1993) 39–59
16. Ziarko, W.: Probabilistic decision tables in the variable precision rough set model. Computational Intelligence **17**(3) (2001) 593–603
17. Yao, Y.Y.: Decision-theoretic rough set models. Lecture Notes in Artificial Intelligence **4481** (2007) 1–12
18. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximating concepts. Int. J. of Man–Machine Studies **37**(6) (1992) 793–809
19. Tsumoto, S.: Modelling medical diagnostic rules based on rough sets. Lecture Notes in Artificial Intelligence **1424** (1998) 475–482
20. Gomolińska, A.: Satisfiability and meaning of formulas and sets of formulas in approximation spaces. Fundamenta Informaticae **67**(1–3) (2005) 77–92
21. Nguyen, H.S., Skowron, A., Stepaniuk, J.: Granular computing: A rough set approach. Computational Intelligence **17**(3) (2001) 514–544
22. Cios, K.J., Pedrycz, W., Swiniarski, R.W., Kurgan, L.A.: Data Mining: A Knowledge Discovery Approach. Springer Science + Business Media, LLC, New York (2007)
23. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. ACM Computing Surveys **31**(3) (1999) 264–323
24. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Chichester (1990)
25. Albatineh, A.N., Niewiadomska-Bugaj, M., Mihalko, D.: On similarity indices and correction for chance agreement. J. of Classification **23** (2006) 301–313
26. Gomolińska, A.: On three closely related rough inclusion functions. Lecture Notes in Artificial Intelligence **4585** (2007) 142–151
27. Gomolińska, A.: On certain rough inclusion functions. Transactions on Rough Sets IX: journal subline of LNCS **5390** (2008) 35–55
28. Gomolińska, A.: Rough approximation based on weak q-RIFs. Transactions on Rough Sets X: journal subline of LNCS **5656** (2009) 117–135
29. Xu, Z.B., Liang, J.Y., Dang, C.Y., Chin, K.S.: Inclusion degree: A perspective on measures for rough set data analysis. Information Sciences **141** (2002) 227–236
30. Zhang, W.X., Leung, Y.: Theory of including degrees and its applications to uncertainty inference. In: Proc. of 1996 Asian Fuzzy System Symposium. (1996) 496–501
31. Borkowski, L., ed.: Jan Łukasiewicz – Selected Works. North Holland/Polish Scientific Publ., Amsterdam/Warsaw (1970)
32. Łukasiewicz, J.: Die logischen Grundlagen der Wahrscheinlichkeitsrechnung, Kraków (1913) English translation in [31], pages 16-63.

33. Drwal, G., Mrózek, A.: System RClass – software implementation of a rough classifier. In Kłopotek, M.A., Michalewicz, M., Raś, Z.W., eds.: Proc. 7th Int. Symp. Intelligent Information Systems (IIS'1998), Malbork, Poland, June 1998. (1998) 392–395
34. Polkowski, L., Tsumoto, S., Lin, T.Y., eds.: Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems. Physica V., Heidelberg New York (2001)