2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)
July 6-11, 2014, Beijing, China

Similarities in Structured Spaces of Sets

Wladyslaw Homenda
Faculty of Mathematics and Information Science,
Warsaw University Technology,
ul. Koszykowa 75, 00-662 Warsaw, Poland
Faculty of Mathematics and Computer Science,
University of Bialystok,
ul. Sosnowa 64, 15-887 Bialystok, Poland,
e-mail: homenda@mini.pw.edu.pl

Agnieszka Jastrzebska
Faculty of Mathematics and Information Science,
Warsaw University Technology,
ul. Koszykowa 75, 00-662 Warsaw, Poland
e-mail: A.Jastrzebska@mini.pw.edu.pl

***Abstract*—The objective of this paper is to present methodology for similarity evaluation of structured spaces of sets inspired by human cognitive processes. In contrast to classical similarity relations, which can operate only within the same space, our method can be applied to separate spaces. Proposed formulas are designed to compare two families of sets belonging to separate spaces. Unlike in set-theoretic approach to similarity present in literature, fundamental knowledge, which we use is sets and subsets cardinalities and division of spaces into subsets combined with appropriate minimum and maximum as aggregation operators. Theoretical discussion is supported with a case study, where we apply designed formulas to calculate similarities of four cities. Introduced method has been constructed after an analysis how humans perform similarity evaluation for hard to compare concepts and phenomena.**

I. Introduction

Similarity is one of prime relations, which allows to infer about complex knowledge and determine dependencies in given environment. For an animate subject, like animals or humans, ability to determine similarities is a basic and extremely common cognitive process, which guarantees correct functioning. Typically, when we investigate similarity of two concepts, we subconsciously imply, that they must have something in common. Such commonalities determine the level of coincidence between two phenomena of interest.

In this article authors present an approach to similarity evaluation of two families of sets from different universes. Since there is no possibility to compare features describing elements from both sets, conventional models fail to assess similarity of two such families. Therefore, contribution discussed in this paper is original and we believe important. In practice, a reliable and scalable method for similarity evaluation of families of sets from distinct spaces is desirable. The objective of our research is to propose similarity relations for two families of sets from disjoint spaces. Our attempts focus on relations based on sets and subsets cardinalities.

The perspective on similarity evaluation discussed in this paper is inspired by human cognitive processes. The researched methodology to similarity evaluation is aimed to describe economic phenomena. Economics is one of the most critical fields, where it is often impossible to apply standard similarity measures to assess the level of coincidence between two sets, for example populations of two countries or cities.

The paper is structured as follows. In Section II we present brief literature review on the topic of similarity. Section III covers methodology for similarity evaluation of sets from distinct spaces. We start with basic formulas, for two sets from distinct spaces and move towards similarity evaluation of structured spaces. Section IV is a case study, where we apply developed methodology to compare cities and their citizens. Section V concludes the paper and points out future research directions.

II. Preliminaries

A. Brief and selected literature review

Similarity in literature has been discussed in various contexts and applications, see for example [6], [7], [9], [10], [11]. We observe huge amount of research on similarity in the area of computer vision, [4], but also psychology and biology. Similarity is also analyzed for imprecise knowledge models: [5], [12].

The topic of similarity is well recognized and often discussed in the literature. Therefore, this review contains only selected highlights. Due to space limitations, we were not able to cover all noteworthy contributions to the area of similarity research.

In the literature the concept of similarity is analyzed conventionally in one of the three main streams:

- Measures of similarity based on distance: Euclidean distance, discrete metrics, Hamming distance and other.
- Other measures of similarity, for example probabilistic-based approach to similarity, which includes correlation coefficients, f-divergence (i.e. Kullback-Leibler divergence), Renyi divergence and others.
- Set-theoretic similarity measures. Some examples are: Dice coefficient, Jaccard index and Tversky index.

We focus on the last approach. Set-theoretic similarity relations discussed in the literature are defined for concepts belonging to the same space. In this sense, this paper contribution is original and new.

Similarity relations based on distance satisfy all metric axioms. They are nonnegative, reflexive, symmetrical and

978-1-4799-2072-3/14/$31.00 ©2014 IEEE

transitive. Distance-based similarity relations are very intuitive in interpretation. Unfortunately they have limited modeling possibilities, as named axioms are very demanding. Especially, when it comes to modeling as abstract terms as concepts and features. Therefore it is necessary to develop versatile similarity relations, applicable for complex objects, especially ones with imprecise knowledge.

The domain of application is fuzzy - features are given with membership degrees. Therefore, metric-based similarity relation in the space of concepts and features may be defined as follows: it is a similarity relation $s : X \times X \to [0,1]$ on a set X, where $s(x,y)$ expresses similarity between x and y satisfying following axioms reflexivity, symmetry and triangle condition $s(x,y) * s(y,z) \leq s(x,z)$, where $*$ denotes a similarity transitivity operator, which in the case of the $[0,1]$ interval, is a t-norm. Let us recall that minimum operator is the most popular t-norm. Depending on the choice of function $*$, in the literature discussed are similarity relations under names of indistinguishability relations, fuzzy equivalence relations, proximity relations and others, [7] [p.5].

An alternative way of constructing similarity measure, which is investigated in this article, is set theoretic approach. Such relations often do not satisfy properties named above. Most prominent set-theoretic methodology for similarity evaluation has been presented by A. Tversky in [13]. We briefly discuss this topic below in Section II-B.

In the set-theoretic approach to similarity modeling of interest are objects belonging to the same space. Similarity between two objects: A and B is obtained by analysis of their descriptions. Descriptions are sets of attributes determining objects' shapes. The more similar are sets A and B, the more common features they share. Existence of features is binary (feature either exists or not).

There are other nonmetric similarity measures and various different approaches to determine dependencies between objects, which are not discussed here, due to space limitations.

B. Tversky's similarity and derived indexes

Let us recall that original Tversky's similarity measure concerns similarity of objects a and b, which are characterized by their sets of features A and B. Similarity of objects a and b is stated in terms of smilarity of their qualitative features. Therefore, having two sets A and B in a space X, $A, B \subset X$, c.f. Figure 1, the Tversky's similarity measure of these sets is expressed by the formula:

$$S_{\alpha,\beta}(A,B) = \frac{|A \cap B|}{|A \cap B| + \alpha \cdot |A \setminus B| + \beta \cdot |B \setminus A|} \qquad (1)$$

where $|\cdot|$ denotes cardinality, $\alpha, \beta \geqslant 0$ are parameters of Tversky's index.

Circles, corresponding to A and B on Figure 1 are more overlapping for more similar objects. If objects are indistinguishable, circles corresponding to them are on each other. Objects, which do not share common features are disjoint. This very basic concept can be generalized onto many domains and can be transferred into plenty different similarity measures.

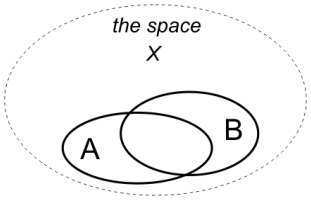

Fig. 1. Tversky's similarity measure.

Tversky's index is asymmetric in general. However, in the case of equality of parameters $\alpha = \beta$ Tversky's index becomes symmetric. The formula can be rewritten as:

$$S_{\alpha,\beta}(A,B) = \\ = \frac{|A \cap B|}{|A \cap B| + \delta \cdot (\gamma \cdot |A \setminus B| + (1-\gamma) \cdot |B \setminus A|)} \qquad (2)$$

which allows to control (a)symmetry as well as balance between intersection of files and their symmetric difference $(A \cup B) \setminus (A \cap B)$. In this formula parameter α keeps control over (a)symmetry while parameter β is responsible for balance between intersection and symmetric difference. Relations between parameters of formulas 1 and 2 are as follows: $\gamma = \alpha/(\alpha+\beta)$ and $\delta = \alpha + \beta$.

Setting $\alpha = 1 = \beta$ gives Tanimoto index, while for $\alpha = 0.5 = \beta$ we get Dice index. By analogy we discuss some issues related to this topic.

III. Similarities

In this section discussion is based on set theoretic Tversky's similarity measure. Alike Tversky's similarity, we consider similarity of sets. Tversky's similarity concerns sets in the same space. Such sets can be subjected to set theoretic operations: union, intersection and complement.

Unlike Tversky's case, we consider files in different spaces. Therefore, set theoretic operations are not applied to files coming from different spaces. Of course, we can formally define intersection of two files of different spaces as the empty set and union of such sets as the set of all elements from both such sets. But this is not the aim and we do not consider such formal operations.

It is worth to highlight again that the presented attempt is motivated and inspired by human cognitive processes.

The section starts with basic notions. First, methodology for similarity evaluation of two sets from distinct spaces is presented. Subsequently, in Subsection III-C, authors discuss similarity of structured spaces of files. We extend the methodology introduced for separate files to account structured spaces.

A. Similarity of files

1) Similarity of separated files: First of all, we consider similarity of two sets which are not comparable, i.e. they are subsets of different spaces (universes). For instance, considering populations of two different cities X and Y, we can compare sets A and B of citizens of both populations. In this way we obtain two sets of individuals, which are not comparable in the sense of Tversky's measure. Indeed,

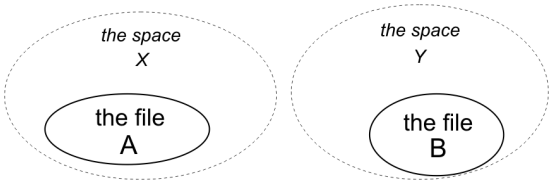

Fig. 2. Similarity of files.

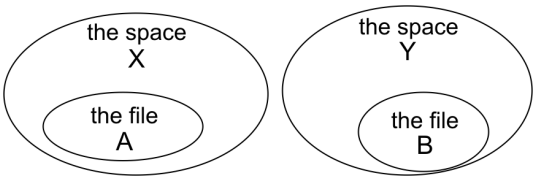

Fig. 3. Similarity of structures: files in spaces.

intersection of such files is empty. Anyway, we still can expect some similarity level of such sets. Intuitively, such two sets are identical, are similar at the level 1, if they are of the same cardinality. On the other hand, empty and nonempty sets are not similar at all.

Let us assume two sets: $A = \{a_1, a_2, \ldots, a_k\}$ and $B = \{b_1, b_2, \ldots, b_l\}$, c.f. Figure 2. As mentioned, elements of both sets are incomparable. Adapting Tversky's similarity measure, we can invent a formula compatible with the above intuitive observations:

$$S(A,B) = \frac{\min\{|A|,|B|\}}{\max\{|A|,|B|\}} = \\ = \frac{\min\{|A|,|B|\}}{\min\{|A|,|B|\} + \big||A|-|B|\big|} = \frac{\min\{k,l\}}{\min\{k,l\} + \big|k-l\big|} \quad (3)$$

what can be rewritten in a form more suitable for further manipulations on differences between files:

$$S(A,B) = \\ = \frac{\min\{|A|,|B|\}}{\min\{|A|,|B|\}+\max\{0,|A|-|B|\}+\max\{0,|B|-|A|\}} \\ = \frac{\min\{k,l\}}{\min\{k,l\} + \max\{0,k-l\} + \max\{0,l-k\}} \quad (4)$$

The last expressions in formulas 3 and 4 involves cardinalities of considered sets. As to this observation we can draw a conclusion, that similarities under discussion may be interpreted as similarities of natural numbers instead of similarities of files. This remark outlines relations between sets and numbers, which is well known in set theory. In this paper we will refer to sets rather then to numbers in order to meet illustrative examples.

B. Asymmetry of similarities

Similarity measures are usually assumed symmetric. This assumption is not always valid. For instance, we say that Bucharest is a little Paris, but we rather do not say that Paris is a little Bucharest. This observation is interpreted that Bucharest is similar to Paris (in some aspects), but we do not consider Paris to be equally similar to Bucharest. Another example: let us consider two consumers such that both have the same needs besides that the second one has extra need to listen to classical music. They are not identical and intuitively, the first one is more similar to the second one than oppositely. In this study, we assume similarity measures to be asymmetrical.

Returning to similarity of two sets A and B, order of them would be important, i.e. the set A may be more similar to the set B than the set B is similar to the set A. This ascertainment is justified by the following another observation (to supplement the above example with consumers): having two sets, the one of lower cardinality is more similar to the one of higher cardinality than oppositely. Introducing two positive factors α and β we can control asymmetry as well as influence of set differences on the similarity. If $\alpha > \beta$ then this intuition is fulfilled, what is expressed in expansion of formula (4):

$$S(A,B) = \\ \frac{\min\{|A|,|B|\}}{\min\{|A|,|B|\}+\alpha\cdot\max\{0,|A|-|B|\}+\beta\cdot\max\{0,|B|-|A|\}} \\ = \frac{\min\{k,l\}}{\min\{k,l\} +\alpha\cdot\max\{0,k-l\} +\beta\cdot\max\{0,l-k\}} \quad (5)$$

C. Similarity of structures

Discussion on similarity of files in section III-A does not conceived spaces, in which those files were included. Now, let us consider not only files alone, but also spaces of their inclusion. Such considerations are motivated, for instance, by an observation that similarities of groups of citizens of two cities depend on relation between populations of these cities. If we wish to compare numbers of owners of cars of a given brand in different cities, populations of these cities should be considered. Otherwise, such comparison would be defective.

1) Files in spaces: Assume that we distinguish subsets $A \subset X$ and $B \subset Y$ of spaces X and Y having in this way, in fact, two structures, c.f. Figure 3. As mentioned above, if both sets A and B have the same cardinality, then they would be considered perfectly similar. However, since these sets are subsets of spaces X and Y, we should consider similarity of structures $A \subset X$ and $B \subset Y$ rather than these files alone. Reasonable is to assume perfect similarity if both sets A and B have the same cardinality and both structures X and Y have the same cardinality. Generalizing this observation we can state that more close are proportions between $|A|$ and $|X|$ and between $|B|$ and $|Y|$, more similar both structures are. The more similar spaces X and Y, the more similar both structures.

Summarizing, let us express these statements as qualified similarity $S(A|X, B|Y)$ of subsets A and B:

$$S(A|X,B|Y) = \frac{\min\left\{\dfrac{|A|}{|X|}, \dfrac{|B|}{|Y|}\right\}}{\max\left\{\dfrac{|A|}{|X|}, \dfrac{|B|}{|Y|}\right\}} \quad (6)$$

Finally, similarity of such structures cannot exceed similarities of spaces X and Y and qualified similarity od sets A and B. This reflection is summarized by the formula:

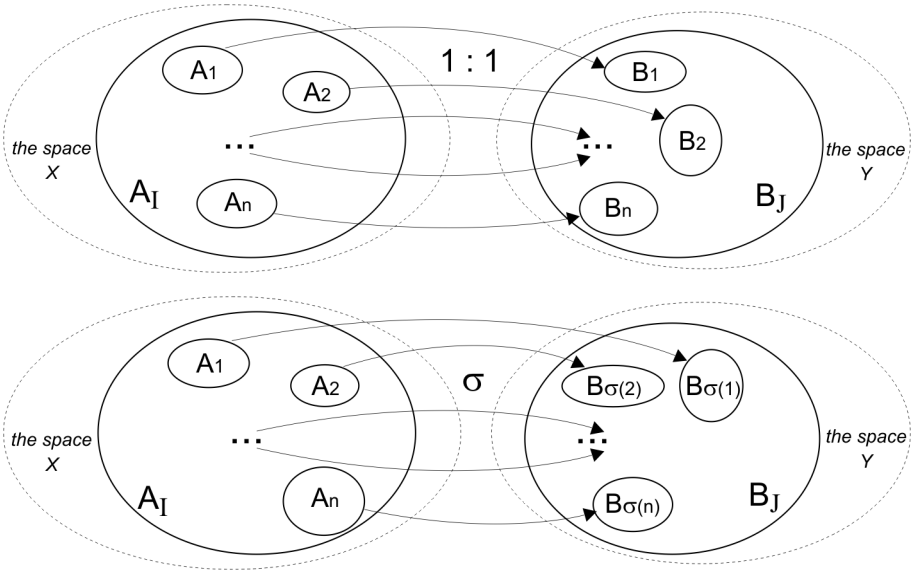

Fig. 4. Qualified similarity of families with identity mapping between families of files (upper part) and with a bijection (permutation of indexes) σ between families of files (bottom part).

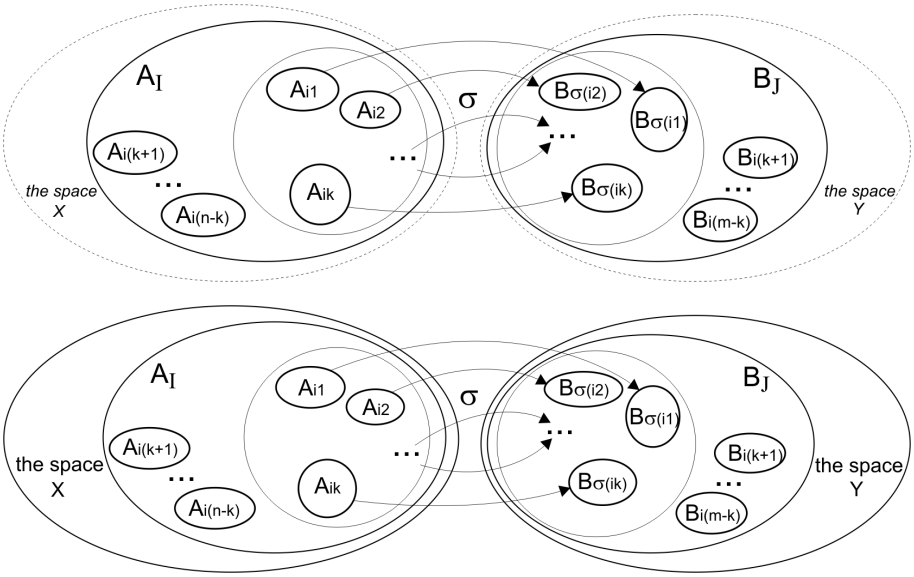

Fig. 5. Similarity of families with optimal subfamilies compared and with admitted permutations of files (upper part) and with the spaces (universes) considered (lower part).

$$S(A \subset X, B \subset Y) = \min\left\{S(X,Y), S(A|X, B|Y)\right\}$$
$$= \min\left\{\frac{\min\{|X|,|Y|\}}{\max\{|X|,|Y|\}}, \frac{\min\left\{\frac{|A|}{|X|}, \frac{|B|}{|Y|}\right\}}{\max\left\{\frac{|A|}{|X|}, \frac{|B|}{|Y|}\right\}}\right\} \quad (7)$$

and the following formula outlines asymmetrical version of similarity of two structures $A \subset X$ and $B \subset Y$:

$$S_{\alpha,\beta}(A \subset X, B \subset Y)$$
$$= \min\left\{S_{\alpha,\beta}(X,Y), S_{\alpha,\beta}(A|X, B|Y)\right\} \quad (8)$$

2) Families of files: In this section the simple structure (a file in a space) is generalized to a family of files in a space. For instance, let us consider populations of two cities. We can distinguish groups of citizens born in a similar time, for instance we can split populations to groups of people born in the same year or groups of people in the same age with 10 years precision etc. In this way we get strict matching of groups of both populations: match two groups (of two populations) if they stand for the same year of birth or have the same age.

Generalizing this observation let us consider two families of sets $\mathcal{A}$ and $\mathcal{B}$. Both families have the same number of sets enumerated by indexes $I = \{1, 2, \ldots, n\}$, i.e. $\mathcal{A} = \{A_1, \ldots, A_n\}$ and $\mathcal{B} = \{B_1, \ldots, B_n\}$. Sets in every

family are pairwise disjoint subsets of spaces X and Y, respectively, i.e. $A_1, \ldots, A_n \subset X$ and $B_1, \ldots, B_n \subset Y$. Denote also $A_I = \cup_{i=1}^n A_i$ and $B_I = \cup_{i=1}^n B_i$. It may happen that files cover spaces, i.e. $A_I = X$ and $B_I = Y$.

At first assume that files match according to their indexes, i.e. that given is identity mapping id between indexes of both families, i.e. correspond files A_i and $B_{id(i)} = B_i$ for $i = 1, 2, \ldots, n$, c.f. upper part of Figure 4. Similarity (qualified, with regard to A_I and B_I) of such families depends on qualified similarity of pairs of corresponding files (qualified similarity of files is described by formula (6)):

$$S_{I,id}(\mathcal{A}|A_I, \mathcal{B}|B_I) = \min_{i \in I} \Big\{ S(A_i|A_I, B_i|B_I) \Big\} \tag{9}$$

In the above formula pairing of files is assumed to be fixed. Usually, comparison might be done much more flexibly. Changing pairing of compared files may give higher similarity than given restrictive one. Assume that correspondence of files is defined by a bijection σ of indexes $I = \{1, 2, \ldots, n\}$ into the same set of indexes I, i.e. corresponding are files A_i and $B_{\sigma(i)}$ for $i = 1, 2, \ldots, n$. As above, similarity of a pair A_i and $B_{\sigma(i)}$ is described by formula (6). Therefore, qualified similarity of families $\mathcal{A}$ and $\mathcal{B}$ with regard to the mapping σ is defined as:

$$S_{I,\sigma}(\mathcal{A}|A_I, \mathcal{B}|B_I) = \min_{i \in I} \Big\{ S(A_i|A_I, B_{\sigma(i)}|B_I) \Big\} \tag{10}$$

Often, no correspondence between files of both families is assumed. In such a case, it is reasonable to find a bijection between sets of indexes, which maximizes similarities between pairs of files, c.f. Figure 4. Then, qualified similarity of families, independent on pairing, is defined as:

$$S_I(\mathcal{A}|A_I, \mathcal{B}|B_I) = \max_{\sigma \in P(I)} S_{I,\sigma}(\mathcal{A}|A_I, \mathcal{B}|B_I) =$$

$$= \max_{\sigma \in P(I)} \Big\{ \min_{i \in I} \Big\{ S(A_i|A_I, B_{\sigma(i)}|B_I) \Big\} \Big\} \tag{11}$$

where $P(I)$ are all bijections between sets of indexes of families (recall that sets of indexes of both families are equal and are equal to I).

In Table I we have $I = \{1, 2, 3, 4\}$ and $\cup_{i=1}^4 A_i = A_I = X$ and $\cup_{i=1}^4 B_i = B_I = Y$. The top part of this tables shows qualified similarities of corresponding files of both families (third row) and qualified similarity of both families for identity mapping (fourth row). The middle part of this Table shows mapping (bijection) σ of files in the family $\mathcal{B}$ (first row), qualified similarities of corresponding files of both families (third row) and qualified similarity of both families for the given mapping (fourth row). Finally, the bottom part (bottom row) gives qualified similarity of both families (maximum of qualified similarities for all bijections), which is equal to 1 in this case.

3) Families of files with subfamilies compared: In this discussion all files were considered in computed similarity of families $\mathcal{A}$ and $\mathcal{B}$. Usually, some files are not important for comparing both families. This gives us an ability to take into account pairs of files of relatively high similarity. But, of course, due to the fact that only parts of families are compared, dropped parts should also be accounted in computation of similarity.

TABLE I. QUALIFIED SIMILARITY OF FAMILES OF FILES, THE SAME CARDINALITY OF FAMILIES, ALL FILES ACCOUNTED.

i	1	2	3	4	I
$\lvert A_i \rvert$	10	20	30	40	100
$\lvert B_i \rvert$	40	30	20	10	100
$S(A_i\|A_I, B_i\|B_I)$	0.25	0.67	0.67	0.25	
$S_{I,id}(\mathcal{A}\|A_I, \mathcal{B}\|B_I)$					0.25
$\sigma(i)$	4	3	2	1	
$\lvert B_{\sigma(i)} \rvert$	10	20	30	40	100
$S(A_i\|A_I, B_{\sigma(i)}\|B_I)$	1	1	1	1	
$S_{I,\sigma}(\mathcal{A}\|A_I, \mathcal{B}\|B_I)$					1
$S_I(\mathcal{A}\|A_I, \mathcal{B}\|B_I)$					1

Assume that given is sequence of (pairwise different) indexes $K = \{i_1, \ldots, i_k\}$ and, of course, $K \subset I = \{1, 2, \ldots, n\}$. Assume also that given is $1 : 1$ mapping (injection) $\sigma : K \to I$. Let us denote $A_K = \cup_{j \in K} A_j$ and $B_K = \cup_{j \in K} B_{\sigma(j)}$. Now, we can define qualified similarity of families $\mathcal{A}$ and $\mathcal{B}$ (with regard to I and σ) based on K as follows:

$$S_{K,\sigma}(\mathcal{A}|A_I, \mathcal{B}|B_I) = \tag{12}$$
$$\min \Big\{ \min_{j \in K} \big\{ S(A_j|A_I, B_{\sigma(j)}|B_I) \big\}, S(A_K \subset A_I, B_K \subset B_I) \Big\}$$

If a mapping σ is not given and can be freely chosen as an injection $\sigma : K \to I$, we get:

$$S_K(\mathcal{A}|A_I, \mathcal{B}|B_I) = \max_{\sigma \in I^K} \Big\{ S_{K,\sigma}(\mathcal{A}|A_I, \mathcal{B}|B_I) \Big\}$$
$$= \max_{\sigma \in I^K} \Big\{ \min \Big\{ S(A_i|A_I, B_{\sigma(i)}|B_I) : i \in K \Big\} \Big\} \tag{13}$$

When no restriction is put on files to be compared, we set the following formula to compute qualified similarity of these two families:

$$S(\mathcal{A}|A_I, \mathcal{B}|B_I) = \max_{K \subset I} \Big\{ S_K(\mathcal{A}|A_I, \mathcal{B}|B_I) \Big\} \tag{14}$$
$$= \max_{K \subset I} \Big\{ \max_{\sigma \in I^K} \Big\{ \min \Big\{ S(A_i|A_I, B_{\sigma(i)}|B_I) : i \in K \Big\} \Big\} \Big\}$$

Considering similarity of unions A_I and B_I in sense of formula (3), we get the following formula:

$$S(\mathcal{A}, \mathcal{B}) = \min\{ S(\mathcal{A}|A_I, \mathcal{B}|B_I), S(A_I, B_I) \} \tag{15}$$

In Table II we have $I = \{1, 2, 3, 4\}$ and $\cup_{i=1}^4 A_i = A_I = X$ and $\cup_{i=1}^4 B_i = B_I = Y$. The top part of this tables shows qualified similarities of corresponding files of both families (third row) and qualified similarity of both families for identity mapping (fourth row). It is easily seen that any permutation of indexes does not change qualified similarity, what directly gives qualified similarity of both families (fifth row).

The bottom part of this Table shows the subfamily K and corresponding qualified similarities for identity mapping on K (second row). Any injection on K does not change this result. Finally, similarity for this given K is shown in third row of this part. Notice, that for this choice of K, similarity

TABLE II. QUALIFIED SIMILARITY OF FAMILES OF FILES, THE SAME CARDINALITY OF FAMILIES, SUBFAMILIES CONSIDERED.

i	1	2	3	4	I
$\|A_i\|$	10	20	30	40	100
$\|B_i\|$	25	25	25	25	100
$S(A_i\|A_I, B_i\|B_I)$	0.4	0.8	0.833	0.625	
$S_{I,id}(\mathcal{A}\|A_I, \mathcal{B}\|B_I)$					0.4
$S_I(\mathcal{A}\|A_I, \mathcal{B}\|B_I)$					0.4
K		2	3	4	
$S(A_K \subset A_I, B_K \subset B_I)$					0.937
$S_{K,id}(\mathcal{A}\|A_I, \mathcal{B}\|B_I)$					0.625
$S_K(\mathcal{A}\|A_I, \mathcal{B}\|B_I)$					0.625
$S(\mathcal{A}\|A_I, \mathcal{B}\|B_I)$					0.625
$S(\mathcal{A}, \mathcal{B})$					0.625

$S(A_K \subset A_I, B_K \subset B_I) = 0.937$ does not diminish the result, c.f. formula 12. It is easy to see that any other choice of K does not increase final similarity, but rather diminishes.

4) Similarity of families of different cardinalities: Assuming that families $\mathcal{A}$ and $\mathcal{B}$ have different numbers of files, say:

- $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$ and $I = \{1, 2 \ldots, n\}$,
- $\mathcal{B} = \{B_1, B_2, \ldots, B_m\}$ and $J = \{1, 2 \ldots, m\}$,

we define $K \subset \{1, 2, \ldots, n\}$ such that cardinality of K does not exceed cardinality of the smaller family: $|K| \leqslant \min(n, m)$, c.f. Figure 5. Then we apply formula (15) for such restricted subset of indexes $K \subset I$.

5) Similarity of families of files in spaces: Finally, if spaces X and Y cannot be left aside, c.f. Figure 5, we replace similarity of files in formula (15) by similarity of files in spaces, coming to the following formula:

$$S(\mathcal{A} \subset 2^X, \mathcal{B} \subset 2^Y) = \min\left\{S(\mathcal{A}, \mathcal{B}), S(A_I \subset X, B_J \subset Y)\right\} \tag{16}$$

Needless to say that computation of similarity of families of sets is a hard optimization problem.

In Table III we have different cardinality of families of files $I = \{1, 2, \ldots, 6\}$ and $J = \{1, 2, \ldots, 10\}$. Alike in former examples, $\cup_{i=1}^{6} A_i = A_I = X$ and $\cup_{j=1}^{10} B_j = B_J = Y$, c.f. top two sections of this Table. Next two sections of the Table show similarities for two different choices of subfamily K and different injections on K. Bottom section (bottom row) shows final similarity of these two families.

Summary of concepts of similarities is outlined in Table III. In this Table we propose names for different kind of similarity measures: direct, qualified and unconditional. Direct similarity measures are applied to pairs of sets. Qualified and unconditional similarity measures are applied to both: pairs of sets and pairs of families.

IV. CASE STUDY

Let us consider a case study example based on 4 Polish cities: Bialystok, Bydgoszcz, Krakow and Warszawa. The aim

TABLE III. QUALIFIED SIMILARITY OF FAMILES, DIFFERENT CARDINALITIES OF FAMILIES, SUBFAMILES CONSIDERED.

i	1	2	3	...	6			I
$\|A_i\|$	50	10	10	...	10			100
j	1	2	3	...	6	...	10	J
$\|B_j\|$	10	10	10	...	10	...	10	100
i	1	2	3	...	6			K
$\|A_i\|$	50	10	10	...	10			100
$\sigma(i)$	1	2	3	...	6			$\sigma(K)$
$\|B_j\|$	10	10	10	...	10			100
$S(A_i\|A_I, B_i\|B_I)$	0.2	1	1	...	1			
$S(A_K \subset A_I, B_K \subset B_I)$								0.5
$S_{K,\sigma}(\mathcal{A}\|A_I, \mathcal{B}\|B_I)$								0.2
i		2	3	...	6			K
$\|A_i\|$		10	10	...	10			100
$\sigma(i)$		2	3	...	6			$\sigma(K)$
$\|B_j\|$		10	10	...	10			100
$S(A_i\|A_I, B_i\|B_I)$		1	1	...	1			
$S(A_K \subset A_I, B_K \subset B_I)$								1
$S_{K,\sigma}(\mathcal{A}\|A_I, \mathcal{B}\|B_I)$								0.4
$S(\mathcal{A}\|A_I, \mathcal{B}\|B_I)$								0.4
$S(\mathcal{A}, \mathcal{B})$								0.4
$S(\mathcal{A} \subset 2^{A_I}, \mathcal{B} \subset 2^{B_I})$								0.4

of this case study is to illustrate properties of the proposed methodology for similarities evaluation of sets and families of sets from distinct structured spaces on real data. We investigate how proposed formulas perform of population data. Attention is on qualified similarity and on similarity of sets' families.

We compare selected cities, one against another. Population of each city is a distinct space. Space structuring is an additional knowledge, which we take benefit of. In this case study spaces structuring is according to populations age. Each space is divided into disjoint subsets consisting of people of certain age. We have distinguished following age groups: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70 years old and over.

We apply formulas introduced in Section III to calculate similarities between aforementioned cities. All details regarding this case study: subsets' sizes and results are gathered in Table V. Numerical information (populations) are from statistical yearbook.

Table's V is organized as follows. In the top section populations of all groups are given and population of cities is given in the last column marked as I. In this case the set of indexes I includes marks of groups: 0-9, 10-19 etc. Notice, that in this case, due to nature of the problem, only identity mapping is considered.

Next three sections are devoted to comparison of three pairs of cities. These pairs are given in the first column. First two rows of each section outlines direct and qualified similarities of corresponding groups of people and, in the last column, direct similarity of whole population of given pair of cities. Then, consecutive four rows of numbers define subfamilies K. Simply, a group of people in certain age is dropped in

TABLE IV. SUMMARY OF SIMILARITIES

	direct	qualified		unconditional
	sets	sets	families	sets and families
$A \subset X$ $B \subset Y$ $X \cap Y = \emptyset$	$S(A,B) = \frac{\min\{\lvert A\rvert, \lvert B\rvert\}}{\max\{\lvert A\rvert, \lvert B\rvert\}}$	$S(A\lvert X, B\lvert Y) = \frac{\min\left\{\frac{\lvert A\rvert}{\lvert X\rvert}, \frac{\lvert B\rvert}{\lvert Y\rvert}\right\}}{\max\left\{\frac{\lvert A\rvert}{\lvert X\rvert}, \frac{\lvert B\rvert}{\lvert Y\rvert}\right\}}$		$S(A \subset X, B \subset Y) = \min\{S(A\lvert X, B\lvert Y), S(X,Y)\}$
$I = \{1,\ldots,i_0\}$ $J = \{1,\ldots,j_0\}$ $\mathcal{A} = \{A_i : i \in I\}$ $\mathcal{B} = \{B_j : j \in J\}$ $A_I = \bigcup_{i\in I} A_i$ $B_J = \bigcup_{j\in J} B_j$ $K \subset I, \lvert K\rvert \leqslant \lvert J\rvert$	$S(A,B) = \frac{\min\{\lvert A_i\rvert, \lvert B_j\rvert\}}{\max\{\lvert A_i\rvert, \lvert B_j\rvert\}}$	$S(A_i\lvert X, B_j\lvert Y) = \frac{\min\left\{\frac{\lvert A_i\rvert}{\lvert X\rvert}, \frac{\lvert B_j\rvert}{\lvert Y\rvert}\right\}}{\max\left\{\frac{\lvert A_i\rvert}{\lvert X\rvert}, \frac{\lvert B_j\rvert}{\lvert Y\rvert}\right\}}$	$S_K(\mathcal{A},\mathcal{B}) = \max_{\sigma\in J^K}\left\{S_{K,\sigma}(\mathcal{A},\mathcal{B})\right\}$ $S(\mathcal{A},\mathcal{B}) = \max_K\left\{S_K(\mathcal{A},\mathcal{B})\right\}$	$S(\mathcal{A}\subset 2^X, \mathcal{B}\subset 2^Y) = \min\left\{S(\mathcal{A},\mathcal{B}), S(A_I \subset X, B_J \subset Y)\right\}$

TABLE V. REAL EXAMPLE

	i	$0-9$	$10-19$	$20-29$	$30-39$	$40-49$	$50-59$	$60-69$	$70\leqslant$	I
Bialystok	$\lvert A_i\rvert/1000$	28.1	28.5	50.5	47.8	38.3	43.8	28.3	29.0	294.3
Bydgoszcz	$\lvert B_i\rvert/1000$	32.4	33.5	55.5	56.9	43.1	56.7	43.7	41.0	363.0
Krakow	$\lvert C_i\rvert/1000$	69.0	62.4	131.0	126.9	88.4	108.5	85.5	87.3	759.1
Warszawa	$\lvert D_i\rvert/1000$	167.8	125.0	250.9	302.3	188.9	254.3	195.6	223.7	1 708.5
	$S(A_i, B_i)$	0.867	0.851	0.909	0.839	0.887	0.773	0.647	0.707	0.811
	$S(A_i\lvert A_I, B_i\lvert B_I)$	0.935	0.953	0.892	0.966	0.914	0.953	0.798	0.872	
		$K \subset I$								
Bialystok	$S_{I,id}(A_i\lvert A_I, B_i\lvert B_I)$	$0-9$	$10-19$	$20-29$	$30-39$	$40-49$	$50-59$	$60-69$	$70\leqslant$	0.798
vs	$S_{K,id}(A_i\lvert A_I, B_i\lvert B_I)$	$0-9$	$10-19$	$20-29$	$30-39$	$40-49$	$50-59$		$70\leqslant$	0.872
Bydgoszcz	$S_{K,id}(A_i\lvert A_I, B_i\lvert B_I)$	$0-9$	$10-19$	$20-29$	$30-39$	$40-49$	$50-59$			0.766
	$S_{K,id}(A_i\lvert A_I, B_i\lvert B_I)$	$0-9$	$10-19$		$30-39$	$40-49$	$50-59$			0.613
	$S_{id}(\mathcal{A}\lvert A_K, \mathcal{B}\lvert B_K)$									0.872
	$S_{id}(\mathcal{A}\subset 2^{A_I}, \mathcal{B}\subset 2^{B_I})$									0.811
	$S(B_i, C_i)$	0.470	0.537	0.424	0.449	0.488	0.523	0.511	0.471	0.478
	$S(B_i\lvert B_I, C_i\lvert C_I)$	0.983	0.890	0.886	0.938	0.981	0.915	0.935	0.984	
		$K \subset I$								
Bydgoszcz	$S_{I,id}(B_i\lvert B_I, C_i\lvert C_I)$	$0-9$	$10-19$	$20-29$	$30-39$	$40-49$	$50-59$	$60-69$	$70\leqslant$	0.886
vs	$S_{K,id}(B_i\lvert B_I, C_i\lvert C_I)$	$0-9$	$10-19$		$30-39$	$40-49$	$50-59$	$60-69$	$70\leqslant$	0.827
Krakow	$S_{K,id}(B_i\lvert B_I, C_i\lvert C_I)$	$0-9$			$30-39$	$40-49$	$50-59$	$60-69$	$70\leqslant$	0.745
	$S_{K,id}(B_i\lvert B_I, C_i\lvert C_I)$	$0-9$			$30-39$	$40-49$		$60-69$	$70\leqslant$	0.600
	$S_{id}(\mathcal{B}\lvert B_K, \mathcal{C}\lvert C_K)$									0.886
	$S_{id}(\mathcal{B}\subset 2^{B_I}, \mathcal{C}\subset 2^{C_I})$									0.478
	$S(C_i, D_i)$	0.411	0.499	0.522	0.420	0.468	0.427	0.437	0.390	0.444
	$S(C_i\lvert C_I, D_i\lvert D_I)$	0.925	0.891	0.851	0.945	0.949	0.961	0.984	0.879	
		$K \subset I$								
Krakow	$S_{I,id}(C_i\lvert C_I, D_i\lvert D_I)$	$0-9$	$10-19$	$20-29$	$30-39$	$40-49$	$50-59$	$60-69$	$70\leqslant$	0.851
vs	$S_{K,id}(C_i\lvert C_I, D_i\lvert D_I)$	$0-9$	$10-19$		$30-39$	$40-49$	$50-59$	$60-69$	$70\leqslant$	0.827
Warszawa	$S_{K,id}(C_i\lvert C_I, D_i\lvert D_I)$	$0-9$	$10-19$		$30-39$	$40-49$	$50-59$	$60-69$		0.712
	$S_{K,id}(C_i\lvert C_I, D_i\lvert D_I)$	$0-9$			$30-39$	$40-49$	$50-59$	$60-69$		0.630
	$S_{id}(\mathcal{C}\lvert C_I, \mathcal{D}\lvert D_I)$									0.851
	$S_{id}(\mathcal{C}\subset 2^{C_I}, \mathcal{D}\subset 2^{D_I})$									0.444

consecutive rows. A group is dropped, for which qualified similarity is the lowest one.

The last but one row of each comparative section gives qualitative similarity of families of files. The last row of each comparative section provides unconditional similarity of discussed people's groups at the background of whole population of given pairs of cities. Notice that there is no variability with regard to mapping between members of both

families. This is why the subscript id is used for each similarity symbol S.

It is worth to draw attention to the firs pair of cities: Bialystok and Bydgoszcz. In this case the highest qualified similarity $S_{K,id}(A_i|A_I, B_i|B_I)$ is reached for the subfamily K with dropped the group of people aged 60-69. For two other pairs the highest similarity is achieve for the whole families I. This is due to diminishing similarity of subfamilies.

Due to space limitation we do not discuss the presented case in details. We also have to resign from comparing provinces of the country. Such comparison leads to computing similarity of families of different cardinalities.

V. Conclusions

The paper introduced methodology for similarity evaluation designed for distinct spaces of sets. Proposed approach provides a wide variety of modeling possibilities. Designed relations allow to calculate similarities of two sets belonging to distinct spaces, and also similarities of two spaces of sets.

To our knowledge, there are no such approaches in the literature. Sets of interest belong to distinct spaces. In such situation we cannot calculate similarity of two distinct spaces using classical methods. Analogically, there are no commonly accepted similarity relations for two distinct spaces. In this context, proposed methodology is an original contribution to the field of widely understood information modeling. The objective of this paper was to introduce key concepts behind the proposed similarity relations.

In contrast to comparison of two concepts belonging to the same space, when we try to compare two concepts from different spaces there are no shared features, that we can assess. Qualitative comparison, which is the crux of the set-theoretic approach to similarity modeling fails for sets belonging to distinct universes and all the more it fails to compare two distinct spaces. Our methodology is based on the only comparable knowledge available in such situations: sets and subsets cardinalities.

Proposed procedure for spaces similarity evaluation takes advantage of underlying division of given two spaces into disjoint subsets. Such structuring is commonly observed in nature - phenomena from the same universe are divided into groups. We take advantage of such division assuming that spaces structuring is a subject of similarity. Presented formulas for similarity calculation are inspired by human cognitive processes and they express rather „pessimistic" evaluations through selected aggregation operators.

The objective of this paper was to introduce researched methodology. We aimed to focus on proposed similarity relations. Theoretical discussion is supported with a case study, where we compared 4 cities by their populations' age.

In future authors plan to continue research in this direction. We are interested in similarities of structured spaces with alternative definitions of the structuring, for example inclusions.

Acknowledgment

The research is supported by the National Science Center, grant No 2011/01/B/ST6/06478, decision no DEC-2011/01/B/ST6/06478.

References

[1] L. Belanche, J. Orozco, *Things to Know about a (dis)similarity Measure*, in: LNAI 6881, 2011, pp. 100-109.

[2] V. D. Blondel, A. Gajardo, M. Heymans, P. Senellart, P. Van Dooren, *A Measure of Similarity between Graph Vertices: Applications to Synonym Extraction and Web Searching* in: SIAM Review, Vol. 46, No. 4, 2004, pp. 647-666.

[3] P.A. Champin, C. Solnon, *Measuring the similarity of labeled graphs* in: Proc. of the Fifth International Conference on Case-Based Reasoning, Springer, 2003, pp. 80-95.

[4] Goshtasby A.A., *Image Registration*, in: Advances in Computer Vision and Pattern Recognition Series, London: Springer, 2012.

[5] Hung W., Yang M., *Similarity measures of intuitionistic fuzzy sets based on Hausdorff distance*, in: Pattern Recognition Letters 25, 2004, pp. 16031611.

[6] Julian-Iranzo P., *A procedure for the construction of a similarity relation*, in: proc. of IPMU'08, 2008, pp. 489 - 496.

[7] Klawonn F., Kruse R., *Similarity Relations and Independence Concepts*, in: G. Della Riccia, D. Dubois, R. Kruse, H.-J. Lenz (eds.): Preferences and Similarities, Springer, Wien, 2008, pp. 179- 196.

[8] G. Li, X. Liu, J. Feng, L. Zhou, *Efficient Similarity Search for Tree-Structured Data*, in: LNCS 5069, 2008, pp. 131-149.

[9] Lin D., *An Information-Theoretic Definition of Similarity*, w: ICML '98 Proceedings, 1998, pp. 296-304.

[10] Orozco J., Belanche L. ,*On Aggregation Operators of Transitive Similarity and Dissimilarity Relations*, w: FUZZ-IEEE, 2004, pp. 1373-1377.

[11] Schoenauer S., *Efficient Similarity Search in Structured Data*, PhD Dissertation defended at the Ludwig-Maximilians-Universitaet Muenchen, 2003

[12] Szmidt E., Kacprzyk J., *A Similarity Measure for Intuitionistic Fuzzy Sets and Its Application in Supporting Medical Diagnostic Reasoning*, in: LNAI 3070, pp. 388393, 2004.

[13] A. Tversky, *Features of Similarity*, Psychological Reviews 84 (4), 1977, pp. 327-352.

[14] R. Yang, P. Kalnis, A.K.H. Tung, *Similarity Evaluation on Tree-structured Data*, in: Proc. of SIGMOD, 2005, pp. 754-765.