2015 IEEE Symposium Series on Computational Intelligence

On Perturbations of Multisets

Maciej Krawczak
Systems Research Institute, Polish Academy of Sciences
Warsaw School of Information Technology
Warsaw, Poland
krawczak@ibspan.waw.pl

Grażyna Szkatuła
Systems Research Institute, Polish Academy of Sciences
Warsaw, Poland
szkatulg@ibspan.waw.pl

Abstract—**The proposed approach is based on the theory of multisets. In the paper we defined a novel measure of remoteness between multisets. There is introduced the definition of perturbation of one multiset by another multiset and vice-versa. In general these measures are different, asymmetrical, so they should not be considered as the distance between considered multisets.**

I. Introduction

Defining a good distance measure between objects is of crucial importance in, for example, many classification and grouping algorithms. From the mathematical point of view, distance is defined as a quantitative degree showing how far apart two objects are. In general, the objects are described by sets of attributes. While a lot of work has been performed on continuous attributes, nominal attributes are more difficult to handle. Nominal data contains data with nominal attributes whose values neither have a natural ordering nor an inherent order. The variables of nominal data are measured by nominal scales. An attribute is nominal if it can take one of a finite number of possible values and, unlike ordinal attributes; these values bear no internal structure. When the attributes are nominal, definitions of the similarity (or dissimilarity) measures become less trivial. Finding similarities between nominal objects by using common distance measures, which are used for processing numerical data, is not applicable here.

Before comparing objects described by nominal attributes we must be equipped with efficient mathematical tools which ensure satisfactory comparisons of two sets of attributes. Here in this paper we focus our attention to multisets. Classical set theory states that a set is a collection of distinct values. If repeating of any values is allowed in a set then such a set is called the *multiset* (sometimes also shortened to *mset* or *bag*). This way, the multiset is understood as a set with additional information about the multiplicity of occurring elements. Let us assume now that every subset of finite set V of nominal values, in which repetition of elements is significant, is called a *multiset*. The first time the term multisets was used by Dedekind in 1898. A complete survey of multisets theory can be found in many papers where several operations and their properties are investigated [1-3, 5-7, 9-12]. A multiset can be expressed using different notations. An exemplary multiset containing one occurrence of *a*, three occurrences of *b* and two occurrence of *c* can be described indifferent ways: $\{a^1, b^3, c^2\}$, $\{a,b,b,b,c,c\}$, $\{a,b,c\}_{1,3,2}$, $\{1/a, 3/b,\ 2/c\}$, $\{a1, b3, c2\}$, or $\{(1,a),(3,b),(2,c)\}$ etc. or use square brackets, depending on adopted notation.

This paper is a continuation as well as extension of authors' previous paper on perturbation of sets [4]. The term "perturbation" is used here in the general sense and should not be confused with that known in mathematics or physics. Here, the term perturbation of a set by another set corresponds to Tversky's [12] considerations about objects' similarities.

Here, instead of the crisp sets we examine multisets and then we introduce an innovative measure of proximity between two multisets. This consideration is based on the theory of the multisets and its basic operations. We introduced *a measure of perturbation of one multiset by another multiset*. This measure defines changes of one multiset after adding another multiset, and vice versa. It is interesting that this measure is not symmetric, it means a value of the measure of perturbation of the first multiset by the second multiset can be different than a value of the measure of perturbation of the second multiset by the first multiset. The measure of perturbation is assumed to return a value from [0, 1], where 1 is interpreted as the most level of perturbation, while 0 is the lowest level of perturbation. The measure is asymmetrical, so it should not be considered as the distance between multisets, but the diversity.

This paper is organized as follows: Section 2 presents preliminaries and basic definitions of the multisets. In Section 3 we present the description of perturbation methodology and the mathematical properties of the measure of perturbation are studied.

II. Preliminaries and the Basic Definitions of Multisets

From a practical point of view multisets are very useful structures arising in many areas of mathematics and computer science. In this section the basic definitions and notions of functions in multisets context are presented.

Let us consider a non-empty and finite set V of nominal values, $V = \{v_1, v_2, ..., v_L\}$, $v_{i+1} \neq v_i$, $\forall i \in \{1, 2, ..., L-1\}$. Let us define a multiset in the following way.

Definition 1. *The multiset S drawn from the ordinary set V can be represented by a set of ordered pairs:*

$$S = \{(k_S(v), v)\},\ \forall v \in V \tag{1}$$

where $k_S : V \to \mathrm{N}=\{0,1,2,\ \dots\}$.

978-1-4799-7560-0/15 $31.00 © 2015 IEEE
DOI 10.1109/SSCI.2015.223

The function $k_S(.)$ is called *a counting function* and the value $k_S(v)$ specifies the number of occurrences of the element $v \in V$ in the multiset *S*. The element which is not included in the multiset *S* has a counting index equal zero.

A multiset *S* (1) drawn from the finite set V can be rewritten as follows

$$S=\{(k_S(v_1),v_1),(k_S(v_2),v_2),\ ...,(k_S(v_L),v_L)\} \qquad (2)$$

where the element $v_1 \in V$ appears $k_S(v_1)$ times, element $v_2 \in V$ appears $k_S(v_2)$ times and so on. Thus the value $k_S(v_i)$, $i = 1,2,...,L$, specifies the number of occurrences of the element $v_i \in V$ in the multiset *S*, where $k_S(v_i) \geq 0$ (note that $v_i \in V$ is ordinary set notation). For simplicity, the elements $v_i \in V$ for $k_S(v_i) = 0$ will be omitted, when it does not lead to confusion in this paper.

The above definition can be easy interpreted by considering the following example.

Example 1. Let us consider multisets drawn from a set *V*, $V = \{a,b,c,d,e,f,g\}$. Exemplary multiset *S* can be described as $S= \{(3,a),(0,b),(0,c),(2,d),(1,e),(5,f),(0,g)\}$. The elements $v \in V$ for $k_S(v) = 0$ may be omitted to simplify the notation and this multiset can be rewritten in a simplified form as $S= \{(3,a),(2,d),(1,e),(5,f)\}$, Fig. 1.

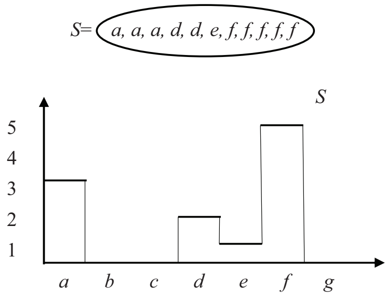

Fig. 1. An exemplary representations of multiset *S* drawn from the ordinary set $V = \{a,b,c,d,e,f,g\}$

□

It should be noted that an ordinary set is a special case of a multiset. Any ordinary subset *A* of the set *V* can be identified by a corresponding multiset $\{(\chi_A(v),v)\}$, $\forall v \in V$, where $\chi_A(.)$ is its characteristic function, $\chi_A : V \to \{0,1\}$. Note, that $\chi_A(v) = 1$ if and only if $v \in A$, i.e., any multiset $S = \{(k_S(v),v)\}$ is an ordinary set if $k_S(v) = 0$ or 1, $\forall v \in V$. In other words, a multiset is a set if a multiplicity of each element is at most one. A multiset *S* is empty, denoted by Ø or {}, if and only if $\forall i \in \{1,2,...,L\}$, $k_S(v_i) = 0$.

Basic notions of multiset are presented below.

The support or *the root* of a multiset *S* drawn from a set *V*, denoted by S^*, is an ordinary set defined as follows: $S^* = \{v \in V :\ k_S(v) > 0\}$ }. Thus, if $\forall v \in V$ such that $k_S(v) > 0$ this implies $v \in S^*$, and $\forall v$ such that $k_S(v) = 0$ this implies $v \notin S^*$. Note that the characteristic function of S^* can be described as $\chi_{S^*}(v) = \min\{k_S(v),1\}$. For example, support of the multiset $S = \{(3,a),(2,d),(5,f)\}$ drawn from a set V can be described as $S^* = \{a,d,f\}$.

The cardinality of the multiset *S*, denoted by $card(S)$ or $|S|$, is defined as the total number of its elements, i.e., $card(S) = \sum_{i=1}^{L} k_S(v_i)$. For example, if $S = \{(3,a),(3,b)\}$ then $card(S) = 6$.

The dimensionality of the multiset *S*, denoted by *dim*(*S*) or $/\, S\, /$, is defined as the total number of various elements, i.e., $dim(S) = \sum_{i=1}^{L} \chi_S(v_i)$. For example, if $S = \{(3,a),(3,b)\}$ then *dim*(*S*)=2.

The multiset space, denoted by $[V]^m$, is the set of all multisets whose elements are in V such that no element occurs more than *m* times.

The complement of the multiset *S* drawn from the set V in the multiset space $[V]^m$ is the multiset S^C such that $k_{S^C}(v) = m - k_S(v)$, $\forall v \in V$. For example, if multiset $S = \{(5,a),(1,b)\}$ drawn from the set $V = \{a,b,c\}$ belongs to the multiset space $[V]^8$ then the complement of its can be described as $S^C = \{(3,a),(7,b),(8,c)\}$.

The upper cut of the multiset *S* drawn from the set V , denoted by $\bar{S}$, is the ordinary set described as $\bar{S} = \{(k_{\bar{S}}(v_1),v_1),(k_{\bar{S}}(v_2),v_2),\ ...,(k_{\bar{S}}(v_L),v_L)\}$,

where $k_{\bar{S}}(v_i) = \begin{cases} 1 & \text{if} \quad k_S(v_i) = 0 \\ 0 & \text{if} \quad k_S(v_i) \geq 1 \end{cases}$, $i \in \{1,2,...,L\}$, $\forall v \in V$.

The upper cut of the multiset *S* drawn from the set V is an ordinary set $\bar{S}$ with elements are not contained in the multiset *S*, but only in the set *V*. For example, for the multiset $S = \{(5,a),(1,b)\}$ drawn from the set $V = \{a,b,c\}$ the upper cut of the multiset *S* can be described as $\bar{S} = \{c\}$.

Example 2. Let us consider the following set $V = \{a,b,c\}$. In Table 1 there are shown a few exemplary functions on multisets S drawn from the set V belonging to the space $[V]^5$.

TABLE 1. FUNCTIONS S^C, S^* AND $\bar{S}$ OF EXEMPLARY MULTISETS

Multiset S	The complement S^C in $[V]^5$	The support S^*	The upper cut $\bar{S}$
$\{(2,b)\}$	$\{(5,a),(3,b),(5,c)\}$	$\{b\}$	$\{a,c\}$
$\{(3,a),(5,b)\}$	$\{(2,a),(5,c)\}$	$\{a,b\}$	$\{c\}$
$\{(2,a),(1,b),(3,c)\}$	$\{(3,a),(4,b),(2,c)\}$	$\{a,b,c\}$	Ø

□

Now, let us consider a finite set V of nominal elements. Assume that S_1 and S_2 are two multisets drawn from the set V, Eq. (2). Below, we present several selected rules of comparison of the multisets.

Equality. We say that two multisets S_1 and S_2 are *equal* or *the same*, denoted by $S_1 = S_2$, if $\forall v \in V$ the condition $k_{S_1}(v) = k_{S_2}(v)$ is satisfied. The following condition holds: if $S_1 = S_2$ then $S_1^* = S_2^*$, however the converse need not holds.

Similarity. We say that two multisets S_1 and S_2 are *similar* if $\forall v \in V$, $k_{S_1}(v) > 0$ if and only if $k_{S_2}(v) > 0$. The similar multisets have equal support sets but need not be equal themselves. For example, multisets $S_1 = \{(3,a),(2,d),(5,f)\}$ and $S_2 = \{(5,a),(1,d),(3,f)\}$ are similar but not equal.

Inclusion. A multiset S_1 is *a sub-multiset* of a multiset S_2, denoted as $S_1 \subseteq S_2$, if $\forall v \in V$ condition $k_{S_1}(v) \le k_{S_2}(v)$ is satisfied.

Many features of operations under multisets are analogues to features of operations under ordinary sets. The basic definitions and notions of the multisets are presented below.

The union of multisets is the multiset defined by

$$\begin{aligned} &S_1 \cup S_2 = \{(k_{S_1 \cup S_2}(v), v): \\ &\forall v \in V, k_{S_1 \cup S_2}(v) = \max\{k_{S_1}(v), k_{S_2}(v)\}\} \end{aligned} \tag{3}$$

For example, if $S_1 = \{(3,a),(3,b)\}$ and $S_2 = \{(5,a),(1,b)\}$ then $S_1 \cup S_2 = \{(5,a),(3,b)\}$.

The intersection of multisets is the multiset defined by

$$\begin{aligned} &S_1 \cap S_2 = \{(k_{S_1 \cup S_2}(v), v): \\ &\forall v \in V, k_{S_1 \cap S_2}(v) = \min\{k_{S_1}(v), k_{S_2}(v)\}\} \end{aligned} \tag{4}$$

For example, if $S_1 = \{(3,a),(3,b)\}$ and $S_2 = \{(5,a),(1,b)\}$ then $S_1 \cap S_2 = \{(3,a),(1,b)\}$.

The arithmetic addition of multisets is the multiset defined by

$$\begin{aligned} &S_1 \oplus S_2 = \{(k_{S_1 \oplus S_2}(v), v): \\ &\forall v \in V,\ k_{S_1 \oplus S_2}(v) = k_{S_1}(v) + k_{S_2}(v)\}. \end{aligned} \tag{5}$$

For example, if $S_1 = \{(3,a),(3,b)\}$ and $S_2 = \{(5,a),(1,b)\}$ then $S_2 \oplus S_1 = \{(8,a),(4,b)\}$.

The arithmetic subtraction of multisets is the multiset defined by

$$\begin{aligned} &S_1 \Theta S_2 = \{(k_{S_1 \Theta S_2}(v), v): \\ &\forall v \in V,\ k_{S_1 \Theta S_2}(v) = \max\{k_{S_1}(v) - k_{S_2}(v), 0\}\} \end{aligned} \tag{6}$$

For example, if $S_1 = \{(3,a),(3,b)\}$ and $S_2 = \{(5,a),(1,b)\}$ then $S_1 \Theta S_2 = \{(0,a),(2,b)\}$ and $S_2 \Theta S_1 = \{(2,a),(0,b)\}$.

The symmetric difference of multisets is the multiset defined by

$$\begin{aligned} &S_1 \Delta S_2 = \{(k_{S_1 \Delta S_2}(v), v): \\ &\forall v \in V,\ k_{S_1 \Delta S_2}(v) = \left|k_{S_1}(v) - k_{S_2}(v)\right|\} \end{aligned} \tag{7}$$

For example, if $S_1 = \{(3,a),(3,b)\}$ and $S_2 = \{(5,a),(1,b)\}$ then $S_1 \Delta S_2 = \{(2,a),(2,b)\}$.

Corollary 1. *Let us assume that we have multisets drawn from the set V. The multisets operations* $\oplus, \cup, \cap$ *satisfy the following properties:*

1. *Commutativity:* $S_1 \oplus S_2 = S_2 \oplus S_1$, $S_1 \cup S_2 = S_2 \cup S_1$
   $S_1 \cap S_2 = S_2 \cap S_1$.
2. *Associativity:* $S_1 \oplus (S_2 \oplus S_3) = (S_1 \oplus S_2) \oplus S_3$
   $S_1 \cap (S_2 \cap S_3) = (S_1 \cup S_2) \cup S_3$
   $S_1 \cap (S_2 \cap S_3) = (S_1 \cap S_2) \cap S_3$.
3. *Idempotence*: $S_1 \cup S_1 = S_1$, $S_1 \cap S_1 = S_1$, $S_1 \oplus S_1 \ne S_1$.
4. *Identity laws*: $S_1 \cup Ø = S_1$, $S_1 \cap Ø = Ø$, $S_1 \oplus Ø = S_1$.
5. *Distributivity*: $S_1 \oplus (S_2 \cup S_3) = (S_1 \oplus S_2) \cup (S_1 \oplus S_3)$
   $S_1 \oplus (S_2 \cap S_3) = (S_1 \oplus S_2) \cap (S_1 \oplus S_3)$
   $S_1 \cup (S_2 \cap S_3) = (S_1 \cup S_2) \cap (S_1 \cup S_3)$
   $S_1 \cap (S_2 \cup S_3) = (S_1 \cap S_2) \cup (S_1 \cap S_3)$.

Assume that we have two multisets with finite support drawn from the set V. It is easy to notice that operator $\oplus$ is stronger than both $\cup$ and $\cap$ in the sense that none of them distributes over $\oplus$, also $(S_1 \cap S_2) \subseteq (S_1 \cup S_2) \subseteq (S_1 \oplus S_2)$ are satisfied. The following property is also satisfied: $card(S_1 \cup S_2) + card(S_1 \cap S_2) = card(S_1) + card(S_2)$.

The above described operations performed on the multisets will be presented by the following example.

Example 3. Let us consider two multisets $S_1 = \{(3,a),(3,b)\}$, $S_2 = \{(5,a),(1,b)\}$, as shown in Fig. 2.

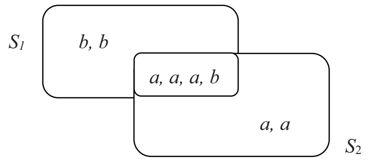

Fig. 2. Exemplary multisets S_1, S_2.

Thus, we obtain the following values of the operations performed on this multisets

$$S_1 \cup S_2 = S_2 \cup S_1 = \{(5,a),(3,b)\}$$
$$S_1 \cap S_2 = S_2 \cap S_1 = \{(3,a),(1,b)\}$$
$$S_1 \oplus S_2 = S_2 \oplus S_1 = \{(8,a),(4,b)\}$$
$$S_1 \Theta S_2 = \{(2,b)\}$$
$$S_2 \Theta S_1 = \{(2,a)\}$$
$$S_1 \Delta S_2 = S_2 \Delta S_1 = \{(2,a),(2,b)\}.$$

□

It should be noted that the sum of the support of the multiset *S* drawn from a set *V* and the upper cut of the multiset *S* gives the set *V*.

Now we will introduce new corollaries which are illustrated by examples.

Corollary 2. *Let us consider the multiset S drawn from the set V. The following property is satisfied*:

$$\bar{S} \oplus S^* = V. \tag{8}$$

Example 4. Let us consider the multiset $S = \{(13,a),(12,d)\}$ drawn from the set *V*={*a, b, c, d*}. Its support can be described as $S^* = \{a,d\}$ and the upper cut as $\bar{S} = \{b,c\}$. The addition of its support and the upper cut provide the entire set *V* and can be written as follows $S^* \oplus \bar{S} = \{a,d\} \oplus \{b,c\} = \{a,b,c,d\}$, thus the formula (8) is obviously satisfied.

□

Corollary 3. *Let us consider the multisets S_1, S_2 drawn from the set V. The following condition $k_{S_1 \cap S_2}(v_i) + k_{S_1 \cup S_2}(v_i) = k_{S_1}(v_i) + k_{S_2}(v_i)$ is satisfied.*

There are several existing methods for comparing the multisets in which distances between multisets are defined on different ways [5, 6, 7].

Instead of considering the similarity or distance measures between two multisets, in the next section, we introduce a new asymmetric measure of remoteness between two multisets.

III. Matching the Multisets

Let us assume that there is a collection of multisets $[V]^m$ drawn from the set *V*, where *V* is a finite set of nominal elements, and $V = \{v_1, v_2, ..., v_L\}$.

Let us consider two multisets S_1 and S_2, such that $S_1, S_2 \subseteq [V]^m$, described by

$$\begin{aligned} S_1 &= \{(k_{S_1}(v_1),v_1),(k_{S_1}(v_2),v_2), ...,(k_{S_1}(v_L),v_L)\}, \\ S_2 &= \{(k_{S_2}(v_1),v_1),(k_{S_2}(v_2),v_2), ...,(k_{S_2}(v_L),v_L)\}. \end{aligned} \tag{9}$$

where counting functions $k_{S_1}(v_i)$ and $k_{S_2}(v_i)$ specify the number of occurrences of the element $v_i \in V$ in the multisets S_1 and S_2, $\forall v_i \in V$, $i \in \{1,2,...,L\}$, $k_{S_1}: V \to \{1,2,...,m\}$, $k_{S_2}: V \to \{0,1,2,...,m\}$. Those elements which are not included in the multisets are accounted as zero.

Let us consider attaching of one multiset set to another. In the case when the first multiset S_1 is attached to the second multiset S_2 we can consider that the second multiset is perturbed by the first multiset or, in other words, the multiset S_1 perturbs the multiset S_2 with some degree. Of course we can consider the counterpart case when the first multiset S_1 is perturbed by the second multiset S_2. In this way we defined a novel *concept of perturbation* of one multiset S_2 by another multiset S_1, which is denoted by $(S_1 \mapsto S_2)$, and interpreted as a multiset $S_1 \Theta S_2$ in the following way

$$\begin{aligned} &(S_1 \mapsto S_2) = S_1 \Theta S_2 = \{(k_{S_1 \mapsto S_2}(v), v): \\ &\forall v \in V,\ k_{S_1 \mapsto S_2}(v) = \max\{k_{S_1}(v) - k_{S_2}(v), 0\}\} \end{aligned} \tag{10}$$

In the case the perturbation of the multiset S_1 by the multiset S_2 the definition of a multiset $(S_2 \mapsto S_1)$ is similar and defined as follows

$$\begin{aligned} &(S_2 \mapsto S_1) = S_2 \Theta S_1 = \{(k_{S_2 \mapsto S_1}(v), v): \\ &\forall v \in V,\ k_{S_2 \mapsto S_1}(v) = \max\{k_{S_2}(v) - k_{S_1}(v), 0\}\} \end{aligned} \tag{11}$$

The geometrical interpretation of the proposed concept of the perturbation of one multiset by another multiset is presented in the following example.

Example 5. Let us consider the following set $V=\{a,b,c,d,e\}$ and an exemplary two multisets $S_1=\{(1,a),(1,e)\}$ and $S_2=\{(1,a),(1,d),(3,e)\}$, $S_1,\ S_2\subseteq[V]^3$. The perturbation of the multiset S_2 by the multiset S_1 is the empty multiset because the following condition ($S_1\mapsto S_2)=S_1\Theta S_2$ =Ø is satisfied. On the other hand, the perturbation of the multiset S_1 by the multiset S_2 is the multiset $(S_2\mapsto S_1)=S_2\Theta S_1=\{(1,d),(2,e)\}$, see Fig. 3.

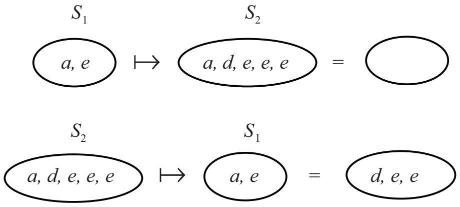

Fig. 3. An exemplary perturbations of the multisets $S_1,S_2\subseteq[V]^3$

□

In order to range the measures of multisets perturbation between 0 and 1 we propose the following way of normalization. Here, we give the following proposal of normalization of the measure of the perturbation of one multiset by another multiset.

Definition 2. *Measure of perturbation of the multiset* S_2 *by the multiset* S_1, *for* $S_1,S_2\subseteq[V]^m$, *denoted by* $Per_{MS}(S_1\mapsto S_2)$ *, is defined in the following manner:*

$$Per_{MS}(S_1\mapsto S_2)=\frac{card(S_1\Theta S_2)}{card(S_1\oplus S_2)}=\frac{\sum_{i=1}^{L}(k_{S_1}(v_i)-k_{S_1\cap S_2}(v_i))}{\sum_{i=1}^{L}(k_{S_1}(v_i)+k_{S_2}(v_i))} \tag{12}$$

In the case the perturbation of the multiset S_1 by the multiset S_2 the definition of a multiset $S_2\Theta S_1$ is similar

$$Per_{MS}(S_2\mapsto S_1)=\frac{card(S_2\Theta S_1)}{card(S_2\oplus S_1)}=\frac{\sum_{i=1}^{L}(k_{S_2}(v_i)-k_{S_2\cap S_1}(v_i))}{\sum_{i=1}^{L}(k_{S_2}(v_i)+k_{S_1}(v_i))} \tag{13}$$

Example 6. Let us consider the set $V=\{a,b,d,e\}$, i.e., L=4, and an exemplary multisets $S_1,S_2\subseteq[V]^4$, where S_1 = $\{(1,a),(1,e)\}$ and S_2 = $\{(1,a),(1,d),(3,e)\}$. Due to Definition 2 we can calculate the following values of the measures of perturbation of one multiset by another, namely:

$$Per_{MS}(S_1\mapsto S_2)=\frac{\sum_{i=1}^{4}(k_{S_1}(v_i)-k_{S_1\cap S_2}(v_i))}{\sum_{i=1}^{4}(k_{S_1}(v_i)+k_{S_2}(v_i))}=\frac{0+0+0+0}{2+0+1+4}=0$$

$$Per_{MS}(S_2\mapsto S_1)=\frac{\sum_{i=1}^{4}(k_{S_2}(v_i)-k_{S_2\cap S_1}(v_i))}{\sum_{i=1}^{4}(k_{S_1}(v_i)+k_{S_2}(v_i))}=\frac{0+0+1+2}{2+0+1+4}=\frac{3}{7}.$$

□

Corollary 4. *Measure of perturbation of the multiset* S_2 *by the multiset* S_1 *satisfies the following condition*

$$0\le Per_{MS}(S_1\mapsto S_2)\le 1\,.$$

Proof.
1) We first prove the first inequality $Per_{MS}(S_1\mapsto S_2)\ \ge 0$. It should be noticed that the inequality $k_{S_1\cap S_2}(v_i)\le k_{S_1}(v_i)$, $\forall i\in\{1,2,...,L\}$ is satisfied, so $k_{S_1}(v_i)-k_{S_1\cap S_2}(v_i)\ge 0$. Using Definition 2 we obtain the following inequality

$$Per_{MS}(S_1\mapsto S_2)=\frac{\sum_{i=1}^{L}(k_{S_1}(v_i)-k_{S_1\cap S_2}(v_i))}{\sum_{i=1}^{L}(k_{S_1}(v_i)+k_{S_2}(v_i))}\ \ge 0\,.$$

2) Let us prove now the second inequality, $Per_{MS}(S_1\mapsto S_2)\le 1$. It should be noticed that the inequality $k_{S_1}(v_i)-k_{S_1\cap S_2}(v_i)\le k_{S_1}(v_i)+k_{S_2}(v_i)$, $\forall i\in\{1,2,...,L\}$ is satisfied. We obtain the following inequality

$$Per_{MS}(S_1\mapsto S_2)=\frac{\sum_{i=1}^{L}(k_{S_1}(v_i)-k_{S_1\cap S_2}(v_i))}{\sum_{i=1}^{L}(k_{S_1}(v_i)+k_{S_2}(v_i))}=$$

$$\le\frac{\sum_{i=1}^{L}(k_{S_1}(v_i)+k_{S_2}(v_i))}{\sum_{i=1}^{L}(k_{S_1}(v_i)+k_{S_2}(v_i))}\ =1.$$

Measure of perturbation of one multiset by another set satisfies the following properties:

Corollary 5. *The following condition is fulfilled* $Per_{MS}(S_1\mapsto S_2)=0$ *if and only if* $k_{S_1}(v_i)=k_{S_1\cap S_2}(v_i)$, $\forall i\in\{1,2,...,L\}$.

Corollary 6. *If* $\forall i\in\{1,2,...,L\}$, $k_{S_2}(v_i)=0$, *and* $\exists k_{S_1}(v_i)>0$, $i\in\{1,2,...,L\}$, *then the condition* $Per_{MS}(S_1\mapsto S_2)=1$ *is satisfied.*

Corollary 7. *The sum of the measures of perturbation of the arbitrary multiset S_2 by another multiset S_1 satisfies the following equality*

$$Per_{MS}(S_1 \mapsto S_2) + Per_{MS}(S_2 \mapsto S_1) = \\ = 1 - \frac{\sum_{i=1}^{L} 2 \cdot k_{S_1 \cap S_2}(v_i))}{\sum_{i=1}^{L} (k_{S_1}(v_i) + k_{S_2}(v_i))} \qquad (14)$$

Definition 3. *Measure of similarity of the multisets S_1 and S_2, $S_1, S_2 \subseteq [V]^m$, denoted by $Sim_{MS}(S_1, S_2)$, is defined in the following manner:*

$$Sim_{MS}(S_1, S_2) = \frac{2\sum_{i=1}^{L} k_{S_1 \cap S_2}(v_i))}{\sum_{i=1}^{L} (k_{S_1}(v_i) + k_{S_2}(v_i))} \qquad (15)$$

Corollary 8. *A measure of similarity is a function which assigns a nonnegative number to every pair of multisets and satisfies the following equations:*

1. $Sim_{MS}(S_1, S_1) = 1$, by Definition 3,
2. $Sim_{MS}(S_1, S_2) = 0$ if and only if $k_{S_1 \cap S_2}(v_i) = 0$, $\forall i \in \{1, 2, ..., L\}$, i.e., $S_1 \cap S_2$ =Ø, by Definition 3,
3. $Sim_{MS}(S_1, S_2) = Sim_{MS}(S_2, S_1)$, by Definition 3,
4. The inequality $0 \le Sim_{MS}(S_1, S_2) \le 1$ is satisfied.

The idea of the measure of similarity described above is illustrated in the following example suggested in the paper's review.

Example 7. Let us consider an example of web mining. We retrieve three texts T_1, T_2, T_3 for some query Q, and count frequencies of four words in each text. Let us denote these four words as: *a*, *b*, *d* and *e*. Respectively, we get the multisets $S_1 = \{(1, a), (1, e)\}$, $S_2 = \{(1, a), (1, d), (3, e)\}$ and $S_3 = \{(1, a), (2, d), (5, e)\}$ drawn from the set $V = \{a, b, d, e\}$, where the numbers show the frequencies of these words in the text T_1, T_2, and T_3, see Fig. 4. The problem is to find the most similar text to the text T_2, i.e., calculate minimum degree of proximity between the multiset S_2 and the multiset S_1 or S_3.

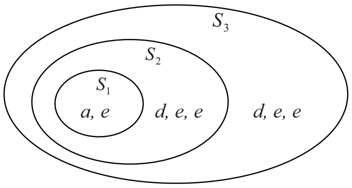

Fig. 4. Exemplary multisets S_1, S_2 and S_3.

Let us note, that S_1 and S_2 differ in 1 for *d* and in 2 in *e*. The differences are the same for S_2 and S_3 for *d* and *e*. For *a* and *b*, there is no difference at all.

Note, that the standard Euclidean distances will be equal, i.e., d(S_1, S_2)=d(S_2, S_3), showing the same similarity.

Let us note, that $card(S_1 \cap S_2)$ = $card(\{(1, a), (1, e)\})$ =2 and $card(S_2 \cap S_3)$ = $card(\{(1, a), (1, d), (3, e)\})$ =5. Thus, the intersection of the multisets S_1 and S_2 has less cardinality than that of the multisets S_2 and S_3. Thus, it seems that they should have not be equally similar.

Let us consider the measures of perturbation of one multiset by another multiset. Due to Definition 2 we can calculate the following values of the measures of perturbation, namely: $Per_{MS}(S_1 \mapsto S_2) = 0$, $Per_{MS}(S_2 \mapsto S_1) = \frac{3}{7}$ and $Per_{MS}(S_2 \mapsto S_3) = 0$, $Per_{MS}(S_3 \mapsto S_2) = \frac{3}{13}$. The measures of perturbation of the multiset S_2 by the multiset S_1 and the multiset S_3 by the multiset S_2 are equal zero. However, the measure of multiset perturbation S_2 by S_3 is less than the measure of multiset perturbation S_1 by S_2 (3/13 vs. 3/7). This way, the multisets S_2 and S_3 are less remoteness than S_2 and S_1.

Now, let us consider the measures of similarity of the multisets. Due to Definition 3 we can calculate the following values of the measures of similarity, namely: $Sim_{MS}(S_1, S_2) = \frac{4}{7}$, $Sim_{MS}(S_2, S_3) = \frac{10}{13}$. Thus, the multisets S_1 and S_2 are less similar than S_2 and S_3 (4/7 vs. 10/13).

Thereby, this short example highlights the differences between the Euclidean distances (which are obviously symmetric) and the multiset perturbations (which are not necessarily symmetric). □

IV. Conclusions

In this paper we propose the new measure of remoteness between two multisets. Some mathematical properties of the measure of perturbation of multisets are explored.

In the authors opinion the methodology presented here is of practical significance. It seems that the approach presented in this paper can be applied to defining a good measure of remoteness between objects. If we defined a description of a group of objects as a *K*-tuple of multisets (i.e., an ordered collection of multisets), we can extend of perturbation on all multisets within describing the considered groups. Instead of considering dissimilarities between groups, we can introduce *a measure of perturbation of one group by another group*. This measure defines changes of one group after adding to another, and vice versa. This methodology needs further research.

Acknowledgment

This work has been partially supported by the National Science Centre Grant UMO-2012/05/B/ST6/03068.

References

[1] A. El-Sayed, Abo-Tabl, Topological approximations of multisets. Journal of the Egyptian Mathematical Society, vol. 21, pp. 123-132, 2013.

[2] K.P. Girish, and J. J. Sunil, Multiset topologies induced by multiset relations. Information Sciences, vol. 188, pp. 298-313, 2012.

[3] D. E. Knuth, The Art of Computer Programming. vol.2. Seminumerical Algorithms, Reading: Addison-Wesley, 1969.

[4] M. Krawczak, and G. Szkatuła, On asymmetric matching between sets. Information Sciences, vol. 312, pp. 89-103, 2015.

[5] A. B. Petrovsky, An Axiomatic Approach to Metrization of Multiset Space. In: Tzeng, G.H., Wang, H.F., Wen, U.P., Yu, P.L., editors, Multiple Criteria Decision Making, New York: Springer-Verlag, pp. 129-1404, 1994.

[6] A. B. Petrovsky, Multiattribute Sorting of Qualitative Objects in Multiset Spaces. In: Koksalan M., Zionts S., editors, Multiple Criteria Decision Making in the New Millennium. Lecture Notes in Economics and Mathematical Systems, N507, Berlin: Springer-Verlag, pp. 124-131, 2001.

[7] A. B. Petrovsky, Cluster Analysis in Multiset Spaces. In: Goldevsky M., Mayr H., editors, Information Systems Technology and its Applications, Bonn: Gesellschaft fur Informatik, pp. 199-206, 2003.

[8] S. Santini, and R. Jain, Similarity Queries in Image Databases, Proceedings of IEEE Conference on Computer vision and Pattern recognition, 1996.

[9] D. Singh, A. M. Ibrahim, T. Yohanna, and J. N. Singh, A systematization of fundamentals of multisets, Lecturas Matematicas, vol. 29, pp. 33-48, 2008.

[10] D. Singh, A. M. Ibrahim, T. Yohanna, and J. N. Singh, An overview of the applications of multisets. Novi Sad J. Math. vol. 37, no. 2, pp. 73-92, 2007.

[11] A. Syropoulos, Mathematics of multisets, In. C.S. Calude at al. (Eds.) Multiset Processing, LNCS 2235, pp. 347-358, 2001.

[12] A. Tversky, Features of similarity, Psychological Review, 84, 327-352, 1977.