ICIT 2015 The 7th International Conference on Information Technology
doi:10.15849/icit.2015.0056 © ICIT 2015 (http://icit.zuj.edu.jo/ICIT15)

A Formal Mathematical Semantics of Advanced Operations of Multiset Table Algebra

Iryna Glushko
Applied Mathematics, Informatics and Educational Measurement Department
Nizhyn Gogol State University
Nizhyn, Ukrain
iryna.glushko@ndu.edu.ua

***Abstract*— Multiset table algebra is considered. The notion of a table specified using the notion of a multiset (or bag). A signature of multiset table algebra is filled up with new operations such as inner and outer joins, semi-join and aggregate operations. A formal mathematical semantics of these operations is defined. The special element NULL is inserted in the universal domain for a define of the outer join.**

Keywords—relation databases; multiset table algebra; inner joins; outer joins; semi-join, aggregate operations

I. Introduction

There are many applications the most peculiar feature of which is multiplicity and repeatability data. For example, these are sociological polls of different population groups, calculations on DNA and others. Commercial relational database systems are almost invariably based on multisets instead of sets. In other words, tables are in general allowed to include duplicate tuples. For example, the data model of SQL is relational in nature, as well as the relevant operations. However, unlike relational algebra, the tables manipulated by SQL are not relations, but, rather, multisets. The reason for this peculiarity is twofold. First, this is due to a practical reason: since SQL tables may be very large, duplicate elimination might become a bottleneck for the computation of the query result. Second, SQL extends the set of query operators by means of aggregate functions, whose operands are in general required to be multisets of values.

The relational model is based on the sets of tuples, i.e. it does not allow duplicate tuples in a relation [1]. So, naturally there is a need to expand possibilities of relational databases due to use of multisets. This problem was also considered in [2-5]. However this question requires specification and extension because in the specified works the due attention isn't paid to operations of inner and outer joins, semijoin and aggregate operations of multiset table algebra.

II. Multiset: Basic Definitions

Let's introduce the basic concepts of multisets in terms of monograph [5].

Definition 1. A multiset α with basis U is a function $\alpha : U \to N^+$, where U is an arbitrary set, $N^+ = \{1,2,\ldots\}$ is the set of natural numbers without zero.

Let D be a universe of element of multiset bases, and then power set $P(D)$ – a universe of multiset bases. Let α be a multiset with basis $U_\alpha = dom\alpha$. Here $dom\alpha$ is the range of definition of multiset as a function.

Definition 2. A characteristic function of multiset α is a function $\chi_\alpha : D \to N$, the values of which are specified by the following piecewise schema: $\chi_\alpha(d) = \begin{cases} \alpha(d) \text{ if } d \in dom\alpha, \\ 0, \text{else;} \end{cases}$ for all $d \in D$.

Definition 3. An empty multiset $\varnothing_m$ is a multiset a characteristic function of which is a constant function, value of which is everywhere equal zero.

Definition 4. A rank of finite multiset α is a sum of duplicate elements of its basis $\|\alpha\| = \sum_{d \in \mathrm{dom}\,\alpha} \alpha(d)$; wherein $\|\varnothing_m\| = 0$.

Let's introduce a binary relation inclusion over multisets.

Definition 5. Multiset β is included in multiset α ($\beta \preceq \alpha$), if $U_\beta \subseteq U_\alpha \,\&\, \forall d \left(d \in U_\beta \Rightarrow \beta(d) \le \alpha(d) \right)$. Directly from definition follows that this relation is a partial order.

The 1-multisets are the multisets whose range of values is an empty set or a single-element set {1}. These multisets are the analogues of ordinary sets.

The operations over multisets are defined in terms of characteristic functions in monograph [5]. There are operations of multiset union $\bigcup_{All}$, intersection $\bigcap_{All}$, difference $\backslash_{All}$, which build multisets of general view. The Cartesian product of multiset $\otimes$, the operation $Dist(\alpha)$, which build 1-multiset, and analog of a full image for multisets are defined too.

III. Multiset Table Algebra: Basic Definitions

Among the two sets that are considered, $\boldsymbol{A}$ is the set of attributes and $\boldsymbol{D}$ is the universal domain.

Definition 6. An arbitrary (finite) set of attributes $R \subseteq \boldsymbol{A}$ is called the scheme.

Definition 7. A tuple of the scheme R is a nominal set on pair R, $\boldsymbol{D}$. The projection of this nominal set for the first component is equal to R.

The set of all tuples on scheme R is designated as $S(R)$ and the set of all tuples is designated as S.

Definition 8. A table is a pair $\langle \psi, R \rangle$, where the first component ψ is an arbitrary multiset basis of which $\Theta(\psi)$ is an arbitrary set (in particular infinite) of tuples of the scheme R and other component R is a scheme of table.

Thus, a certain scheme is ascribed to every table. The set of all table on scheme R is designated as $\Psi(R)$ and the set of all table is designated as $\Psi = \bigcup\limits_{R} \Psi(R)$.

The notation $Occ(s,\psi)$ denotes the number of duplicate tuple s in the multiset ψ. Let's agree a multiset to write down as $\{s_1^{n_1},\ldots,s_k^{n_k}\}$, where $n_i = Occ(s_i,\psi)$, $i = 1,\ldots,k$, and $\Theta(\psi) = \{s_1,\ldots,s_k\}$ is a basis of the multiset ψ.

Definition 9. The multiset table algebra is the algebra $\langle \Psi, \Omega_{P,\Xi} \rangle$, where Ψ is the set of all tables, $\Omega_{P,\Xi} = \{\bigcup_{All}^{R}, \bigcap_{All}^{R}, \backslash_{All}^{R}, \sigma_{p,R}, \pi_{X,R}, \underset{R_1,R_2}{\otimes}, Rt_{\xi,R}, \sim_R\}_{X,R,R_1,R_2 \subseteq \boldsymbol{A}}^{p\in P,\xi\in\Xi}$ is the signature, P, Ξ are the sets of parameters.

The operations of signature $\Omega_{P,\Xi}$ are defined in [6].

IV. The Advanced Operations

The advanced operations include inner and outer joins, semijoin, aggregate operations.

A. Inner Joins

There are four kinds of inner join operations Cartesian join, natural join, join using attributes $A_1,\ldots,A_n$ and join on predicate p. Let's define them.

Definition 10. The Cartesian Join of table on scheme R_1 and table on scheme R_2, moreover $R_1 \cap R_2 = \varnothing$, is a binary parametric operation $\underset{R_1,R_2}{Cj}$ of the form $\underset{R_1,R_2}{Cj} : \Psi(R_1) \times \Psi(R_2) \to \Psi(R_1 \cup R_2)$, $\langle \psi_1, R_1 \rangle \underset{R_1,R_2}{Cj} \langle \psi_2, R_2 \rangle = \langle \psi', R_1 \cup R_2 \rangle$, where $\langle \psi_1, R_1 \rangle \in \Psi(R_1)$, $\langle \psi_2, R_2 \rangle \in \Psi(R_2)$.

The basis of the multiset ψ' is defined by follow: $\Theta(\psi') = \{s \mid \exists s_1 \exists s_2 (s_1 \in \Theta(\psi_1) \wedge s_2 \in \Theta(\psi_2) \wedge s = s_1 \cup s_2)\}$. The number of duplicates is given by the following formula: $Occ(s,\psi') = Occ(s_1,\psi_1) \cdot Occ\,(s_2,\psi_2)$, where $s \in \Theta(\psi')$ and $s = s_1 \cup s_2$.

Definition 11. The Inner Natural Join of table on scheme R_1 and table on scheme R_2 is a binary parametric operation written as $\underset{R_1,R_2}{\otimes}$, whose value is the table on scheme $R_1 \cup R_2$ consisting of all the unions of compatible tuples of input tables. Hence, $\underset{R_1,R_2}{\otimes} : \Psi(R_1) \times \Psi(R_2) \to \Psi(R_1 \cup R_2)$, $\langle \psi_1, R_1 \rangle \underset{R_1,R_2}{\otimes} \langle \psi_2, R_2 \rangle = \langle \psi', R_1 \cup R_2 \rangle$, where $\psi_1 \in \Psi(R_1)$, $\psi_2 \in \Psi(R_2)$.

In other words, each tuple of ψ_1 is paired with each tuple of ψ_2, regardless of whether it is a duplicate or not. The basis of the multiset ψ' is defined by follow: $\Theta(\psi') = \{s \mid \exists s_1 \exists s_2 (s_1 \in \Theta(\psi_1) \wedge s_2 \in \Theta(\psi_2) \wedge s_1 \approx s_2 \wedge s = s_1 \cup s_2)\}$. The number of duplicates is given by the following formula: $Occ(s_1 \cup s_2,\psi') = Occ(s_1,\psi_1) \cdot Occ\,(s_2,\psi_2)$, where $s' \in \Theta(\psi')$ and $s' = s_1 \cup s_2$. The relation $\approx$ is a binary relation of compatibility of tuples $s_1 \approx s_2 \overset{dif}{\Leftrightarrow} s_1 \mid R = s_2 \mid R$ and $s_i \mid R$ is the restrictions of tuple s_i on the scheme R [5].

The Inner Join using $A_1,\ldots,A_n$ of table on scheme R_1 and table on scheme R_2, moreover $R_1 \cap R_2 = \{A_1,\ldots,A_n\}$, is a binary parametric operation of the form $\underset{A_1,\ldots,A_n,R_1,R_2}{\otimes} : \Psi(R_1) \times \Psi(R_2) \to \Psi(R_1 \cup R_2)$, $\langle \psi_1, R_1 \rangle \underset{A_1,\ldots,A_n,R_1,R_2}{\otimes} \langle \psi_2, R_2 \rangle = \langle \psi', R_1 \cup R_2 \rangle$, where $\langle \psi_1, R_1 \rangle \in \Psi(R_1)$, $\langle \psi_2, R_2 \rangle \in \Psi(R_2)$.

Moreover all $A_1,\ldots,A_n$ are pairwise different, $n \geq 1$, and $R_1 \cap R_2 = \{A_1,\ldots,A_n\}$. If input tables have also other general attributes which differ from $A_1,\ldots,A_n$, before join they needs to be renamed.

The basis of the multiset ψ' is defined by follow: $\Theta(\psi') = \{s \mid \exists s_1 \exists s_2 (s_1 \in \Theta(\psi_1) \wedge s_2 \in \Theta(\psi_2) \wedge$

doi:10.15849/icit.2015.0056 © ICIT 2015 (http://icit.zuj.edu.jo/ICIT15)

$\wedge \bigwedge_{i=1}^{n} s_1(A_i) = s_2(A_i) \wedge s = s_1 \cup s_2)\}$. The number of duplicates is given by the following formula: $Occ(s,\psi') = Occ(s_1,\psi_1) \cdot Occ\,(s_2,\psi_2)$, where $s \in \Theta(\psi')$ and $s = s_1 \cup s_2$.

Let $p : S \times S \overset{\sim}{\to} \{true,\ false\}$ be a partial binary predicate on the set of all tuples S such that $\forall s_1 \forall s_2 (\langle s_1, s_2 \rangle \in \operatorname{dom} p \wedge p(s_1, s_2) = true \Rightarrow s_1 \approx s_2)$.

Definition 12. The Inner Join on predicate p of table on scheme R_1 and table on scheme R_2 is a binary partial parametric operation of the form $\underset{p,R_1,R_2}{\otimes} : \Psi(R_1) \times \Psi(R_2) \overset{\sim}{\to} \Psi(R_1 \cup R_2)$, $\langle \psi_1, R_1 \rangle \underset{p,R_1,R_2}{\otimes} \langle \psi_2, R_2 \rangle = \langle \psi', R_1 \cup R_2 \rangle$.

The range of definition of this operation is $\operatorname{dom} \underset{p,R_1,R_2}{\otimes} = \{\langle \langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle \rangle \mid \Theta(\psi_1) \times \Theta(\psi_2) \subseteq \operatorname{dom} p\}$. The basis of the multiset ψ' is defined by follow: $\Theta(\psi') = \{s \mid \exists s_1 \exists s_2 (s_1 \in \Theta(\psi_1) \wedge s_2 \in \Theta(\psi_2) \wedge p(s_1, s_2) \simeq true \wedge$ $\wedge s = s_1 \cup s_2)\}$ and $\simeq$ is a generalized equality (strong Kleene's equality) [7]. The number of duplicates is given by the following formula: $Occ(s_1 \cup s_2, \psi') = Occ(s_1, \psi_1) \cdot Occ\,(s_2, \psi_2)$, where $s' \in \Theta(\psi')$ and $s' = s_1 \cup s_2$.

Let's note the following obvious fact. The join $\underset{R_1,R_2}{\otimes}$ is extension of another arbitrary inner join operation in the following sense:

$$\langle t_1, R_1 \rangle \underset{R_1,R_2}{Cj} \langle t_2, R_2 \rangle = \langle t_1, R_1 \rangle \underset{R_1,R_2}{\otimes} \langle t_2, R_2 \rangle,$$

$$\langle t_1, R_1 \rangle \underset{A_1,\ldots,A_n,R_1,R_2}{\otimes} \langle t_2, R_2 \rangle = \langle t_1, R_1 \rangle \underset{R_1,R_2}{\otimes} \langle t_2, R_2 \rangle,$$

$$\left(\langle t_1, R_1 \rangle \underset{p,R_1,R_2}{\otimes} \langle t_2, R_2 \rangle\right)_1 \subseteq \left(\langle t_1, R_1 \rangle \underset{R_1,R_2}{\otimes} \langle t_2, R_2 \rangle\right)_1 [a].$$

The values of these operations in the left parts of these two equalities and inclusion must be defined.

Definition 13.The Semi-join of table on scheme R_1 and table on scheme R_2 is a binary parametric operation written as $\ltimes_{R_1,R_2}$, whose value is the table on scheme R_1 containing tuples of the first table which are included in the inner natural join of input tables. Thus, $\ltimes_{R_1,R_2} : \Psi(R_1) \times \Psi(R_2) \to \Psi(R_1)$, $\langle \psi_1, R_1 \rangle \ltimes_{R_1,R_2} \langle \psi_2, R_2 \rangle = \langle \psi', R_1 \rangle$, where $\langle \psi_1, R_1 \rangle \in \Psi(R_1)$, $\langle \psi_2, R_2 \rangle \in \Psi(R_2)$. The basis of the multiset ψ' is defined by follow: $\Theta(\psi') = \{s_1 \mid s_1 \in \Theta(\psi_1) \wedge \exists s_2 (s_2 \in \Theta(\psi_2) \wedge s_1 \approx s_2)\}$. The number of duplicates is given by the following formula: $Occ(s, \psi') = Occ(s, \psi_1)$, where $s \in \Theta(\psi')$.

B. Outer Join

We can lose information when using the inner join operations because the tuples which are not compatible will not be represented in the output table. The outer join operations use when it is necessary to consider the tuples of input tables which didn't get to result of the inner join operations.

Let $NULL$ be a special element of the universal domain $\boldsymbol{D}$. $NULL$ used to denote absent values in the output table. Let $s_{R,NULL}$ be a constant tuple on scheme R, i.e. $s_{R,NULL} : R \to \{NULL\}$.

There is one logical scheme for definition of the outer join operations [5].

Let $\varphi : \Psi(R_1) \times \Psi(R_2) \overset{\sim}{\to} \Psi(R_1 \cup R_2)$ be some partial binary operation on the set of all tables and $(\varphi(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle))_1 \preceq \left(\langle \psi_1, R_1 \rangle \underset{R_1,R_2}{\otimes} \langle \psi_2, R_2 \rangle\right)_1$ for all tables $\langle \langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle \rangle \in \operatorname{dom} \varphi$.

Let's notice that the operations $\underset{R_1,R_2}{Cj}$, $\underset{R_1,R_2}{\otimes}$, $\underset{A_1,\ldots,A_n,R_1,R_2}{\otimes}$, $\underset{p,R_1,R_2}{\otimes}$ are such.

We fix two tables $\langle \psi_1, R_1 \rangle$, $\langle \psi_2, R_2 \rangle$ from range of definition of the operation φ.

Then the table $\langle \psi_1, R_1 \rangle$ takes the following form $\left\langle \psi_1, R_1 \right\rangle = \left\langle \psi_1 \underset{\varphi}{\cap} \psi_2, R_1 \right\rangle \cup_{All}^{R_1} \left\langle \psi_1 \underset{\varphi}{-} \psi_2, R_1 \right\rangle$.

Consider the table $\left\langle \psi_1 \underset{\varphi}{\cap} \psi_2, R_1 \right\rangle = \langle \psi', R_1 \rangle$. The basis of the multiset ψ' is defined by follow: $\Theta(\psi') = \{s_1 \mid s_1 \in \Theta(\psi_1) \wedge \exists s_2 (s_2 \in \Theta(\psi_2) \wedge$ $\wedge s_1 \cup s_2 \in \Theta((\varphi(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle))_1))\}$. The number of duplicates is given by the following formula: $Occ(s_1, \psi') = Occ(s_1, \psi_1)$, where $s_1 \in \Theta(\psi')$.

Consider the table $\left\langle \psi_1 \underset{\varphi}{-} \psi_2, R_1 \right\rangle = \langle \psi'', R_1 \rangle$. The basis of the multiset ψ'' is defined by follow: $\Theta(\psi'') = \{s_1 \mid s_1 \in \Theta(\psi_1) \wedge \forall s_2 (s_2 \in \Theta(\psi_2) \Rightarrow$ $\Rightarrow s_1 \cup s_2 \notin \Theta((\varphi(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle))_1))\}$. The number of duplicates is given by the following formula: $Occ(s_1, \psi'') = Occ(s_1, \psi_1)$, where $s_1 \in \Theta(\psi'')$.

[a] $(\langle t, R \rangle)_1$ is the first component of the pair $\langle t, R \rangle$, i.e., is the set t.

In other words, the tuples of the table $\left\langle \psi_1 \underset{\varphi}{\cap} \psi_2, R_1 \right\rangle$ are used in formation of result of the join operation, and tuples of the table $\left\langle \psi_1 \underset{\varphi}{-} \psi_2, R_1 \right\rangle$ are not used.

We obtain a representation of the table $\langle \psi_2, R_2 \rangle$ replacing the roles of the tables $\langle \psi_1, R_1 \rangle$ and $\langle \psi_2, R_2 \rangle$ in the presentation of the table $\langle \psi_1, R_1 \rangle$.

Let's notice that if the operation φ coincides with the operation $\underset{R_1,R_2}{\otimes}$ then the table $\left\langle \psi_1 \underset{\varphi}{\cap} \psi_2, R_1 \right\rangle$ is the semi-join of the tables $\langle \psi_1, R_1 \rangle$ and $\langle \psi_2, R_2 \rangle$, i.e. $\left\langle \psi_1 \underset{\varphi}{\cap} \psi_2, R_1 \right\rangle = \langle \psi_1, R_1 \rangle \ltimes_{R_1,R_2} \langle \psi_2, R_2 \rangle$.

There are four kinds of the outer joins operations which are induced of the inner join operation φ: outer left join, outer right join, outer full join and union join. Let’s define them.

Consider the following inner natural joins $\left\langle \psi_1 \underset{\varphi}{-} \psi_2, R_1 \right\rangle \underset{R_1,R_2\setminus R_1}{\otimes} \left\langle \{s^1_{R_2\setminus R_1,NULL}\}, R_2 \setminus R_1 \right\rangle = \langle \psi', R_1 \cup R_2 \rangle$, where $\Theta(\psi') = \{s_1 \cup s^1_{R_2\setminus R_1,NULL} \mid s_1 \in \Theta(\psi_1 \underset{\varphi}{-} \psi_2)\}$, $Occ(s',\psi') = Occ(s_1, \psi_1 \underset{\varphi}{-} \psi_2)$, $s' \in \Theta(\psi')$, $s' = s_1 \cup s^1_{R_2\setminus R_1,NULL}$ and $\left\langle \psi_2 \underset{\varphi}{-} \psi_1, R_2 \right\rangle \underset{R_2,R_1\setminus R_2}{\otimes} \left\langle \{s^1_{R_1\setminus R_2,NULL}\}, R_1 \setminus R_2 \right\rangle = \langle \psi'', R_1 \cup R_2 \rangle$, where $\Theta(\psi'') = \{s^1_{R_1\setminus R_2,NULL} \cup s_2 \mid s_2 \in \Theta(\psi_2 \underset{\varphi}{-} \psi_1)\}$, $Occ(s'',\psi'') = Occ(s_2, \psi_2 \underset{\varphi}{-} \psi_1)$, $s'' \in \Theta(\psi'')$, $s'' = s^1_{R_1\setminus R_2,NULL} \cup s_2$.

Definition 14. The Outer Left Join operation is a partial binary operation of the form $\varphi_l : \Psi(R_1) \times \Psi(R_2) \overset{\sim}{\to} \Psi(R_1 \cup R_2)$, where $\operatorname{dom} \varphi_l = \operatorname{dom} \varphi$ and

$$\varphi_l\left(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle\right) = \varphi\left(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle\right) \cup^{R_1\cup R_2}_{All}$$
$$\cup^{R_1\cup R_2}_{All} \left\langle \psi_1 \underset{\varphi}{-} \psi_2, R_1 \right\rangle \underset{R_1,R_2\setminus R_1}{\otimes} \left\langle \{s^1_{R_2\setminus R_1,NULL}\}, R_2 \setminus R_1 \right\rangle.$$

Definition 15. The Outer Right Join operation is a partial binary operation of the form $\varphi_r : \Psi(R_1) \times \Psi(R_2) \overset{\sim}{\to} \Psi(R_1 \cup R_2)$, where $\operatorname{dom} \varphi_r = \operatorname{dom} \varphi$ and

$$\varphi_r\left(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle\right) = \varphi\left(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle\right) \cup^{R_1\cup R_2}_{All}$$
$$\cup^{R_1\cup R_2}_{All} \left\langle \psi_2 \underset{\varphi}{-} \psi_1, R_2 \right\rangle \underset{R_2,R_1\setminus R_2}{\otimes} \left\langle \{s^1_{R_1\setminus R_2,NULL}\}, R_1 \setminus R_2 \right\rangle.$$

Definition 16. The Outer Full Join operation is a partial binary operation of the form $\varphi_f : \Psi(R_1) \times \Psi(R_2) \overset{\sim}{\to} \Psi(R_1 \cup R_2)$, where $\operatorname{dom} \varphi_f = \operatorname{dom} \varphi$ and

$$\varphi_f\left(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle\right) = \varphi\left(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle\right) \cup^{R_1\cup R_2}_{All}$$
$$\cup^{R_1\cup R_2}_{All} \left\langle \psi_1 \underset{\varphi}{-} \psi_2, R_1 \right\rangle \underset{R_1,R_2\setminus R_1}{\otimes} \left\langle \{s^1_{R_2\setminus R_1,NULL}\}, R_2 \setminus R_1 \right\rangle \cup^{R_1\cup R_2}_{All}$$
$$\cup^{R_1\cup R_2}_{All} \left\langle \psi_2 \underset{\varphi}{-} \psi_1, R_2 \right\rangle \underset{R_2,R_1\setminus R_2}{\otimes} \left\langle \{s^1_{R_1\setminus R_2,NULL}\}, R_1 \setminus R_2 \right\rangle.$$

Definition 17. The Outer Union Join operation is a partial binary operation of the form $\varphi_{\cup} : \Psi(R_1) \times \Psi(R_2) \overset{\sim}{\to} \Psi(R_1 \cup R_2)$, where $\operatorname{dom} \varphi_{\cup} = \operatorname{dom} \varphi$ and

$$\varphi_{\cup}\left(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle\right) =$$
$$= \left\langle \psi_1 \underset{\varphi}{-} \psi_2, R_1 \right\rangle \underset{R_1,R_2\setminus R_1}{\otimes} \left\langle \{s^1_{R_2\setminus R_1,NULL}\}, R_2 \setminus R_1 \right\rangle \cup^{R_1\cup R_2}_{All}$$
$$\cup^{R_1\cup R_2}_{All} \left\langle \psi_2 \underset{\varphi}{-} \psi_1, R_2 \right\rangle \underset{R_2,R_1\setminus R_2}{\otimes} \left\langle \{s^1_{R_1\setminus R_2,NULL}\}, R_1 \setminus R_2 \right\rangle.$$

C. Aggregate Operations

The five types of aggregate operations discussed in this article are SUM, AVERAGE, MAXIMUM, MINIMUM, COUNT. The aggregate operations transform a finite table into a table with single tuple and single attribute.

Consider the table $\langle \psi, R \rangle \in \Psi(R)$, where ψ is a finite multiset and $A \in R$. Let α_A be a multiset of column with attribute A of table $\langle \psi, R \rangle$ which contains all elements including duplicates.

Then $\Theta(\alpha_A) = \{d \mid \exists s (s \in \Theta(\psi) \wedge \langle A, d \rangle \in s)\} = \{d \mid \{\langle A, d \rangle\} \in \Theta\left(\left(\pi_{\{A\},R}\left(\langle \psi, R \rangle\right)\right)_1\right)\}$ is an analogue of active domain of the attribute A [5]. The number of duplicates of element $d \in \Theta(\alpha_A)$ is given by the following formula:

$$\alpha_A(d) = Occ(\{\langle A, d \rangle\}, \left(\pi_{\{A\},R}\left(\langle \psi, R \rangle\right)\right)_1) = \sum_{\substack{s \in \Theta(\psi), \\ s(A)=d}} Occ(s, \psi).$$

Let $2^{\boldsymbol{D}'}_m = \{\alpha \mid \Theta(\alpha_A) \in 2^{\boldsymbol{D}'}$ be a family of all multisets, bases of which are the finite subsets of the set $\boldsymbol{D}'$. Here $\boldsymbol{D}' \subseteq \boldsymbol{D}$ is a subset of the universal domain.

Let Num is a numerical subset of the universal domain $\boldsymbol{D}$ that is closed under addition. Extend the set Num by the special element *NULL*. We will not extend the operation of addition to the case where at least one of the arguments is *NULL*.

Let’s define the aggregate operations. At first the five aggregate functions – count, sum, average, maximum, minimum – are defined on a finite multiset and then these functions are transferred to the tables.

Definition 18. The aggregate operation $Sum_{A,R}$ by the attribute A of the finite table on scheme R, $A \in R$, is a unary parametric operation of the form $Sum_{A,R} : \Psi(R) \to \Psi(\{A\})$,

$Sum_{A,R}(\langle\psi,R\rangle)=\left\langle\left\{\left\{\langle A,Sum(\alpha_A)\rangle\right\}^1\right\},\{A\}\right\rangle$[b], where $\langle\psi,R\rangle\in\Psi(R)$. The $Sum(\alpha_A)$ function is applied to a column with attribute A in the table $\langle\psi,R\rangle$, the result obtained is the sum of every value occurrence in α_A. In addition, $NULL$ values don't undertake in attention and it is assumed that the column contains only data of numeric type.

Thus, $Sum:2_m^{Num}\to Num$,

$$Sum(\alpha_A)=\begin{cases}NULL\ if\ \Theta(\alpha_A)=\varnothing;\\ NULL\ if\ \Theta(\alpha_A)=\{NULL\};\\ \sum\limits_{d\in\Theta(\alpha_A)\setminus\{NULL\}}d\alpha_A(d)\ if\ \Theta(\alpha_A)\setminus\{NULL\}\neq\varnothing.\end{cases}$$

So, we have $Sum(\{NULL^n\})=NULL$, $Sum(\{d_1^{n_1},\dots,d_k^{n_k}\})=\sum\limits_{i=1}^{k}d_in_i$ if all elements d_i, $i=\overline{1,k}$, differ from $NULL$.

In the case of the empty table $\langle\psi_\varnothing,R\rangle$ we have $Sum_{A,R}(\langle\psi_\varnothing,R\rangle)=\left\langle\left\{\left\{\langle A,NULL\rangle\right\}^1\right\},\{A\}\right\rangle$, here $\psi_\varnothing=\varnothing_m$.

Example 1. Let $\langle\psi,R\rangle$ be the table of Fig 1. Then $Sum_{A,R}(\langle\psi,R\rangle)=\left\langle\left\{\left\{\langle A,8\rangle\right\}^1\right\},\{A\}\right\rangle$, $Sum_{B,R}(\langle\psi,R\rangle)=\left\langle\left\{\left\{\langle B,6\rangle\right\}^1\right\},\{B\}\right\rangle$, $Sum_{C,R}(\langle\psi,R\rangle)=\left\langle\left\{\left\{\langle C,6\rangle\right\}^1\right\},\{C\}\right\rangle$. In Fig. 2, Fig. 3 and Fig. 4 reactively, we see the tables $Sum_{A,R}(\langle\psi,R\rangle)$, $Sum_{B,R}(\langle\psi,R\rangle)$, $Sum_{C,R}(\langle\psi,R\rangle)$.

A	B	C
NULL	0	3
2	1	1
2	1	1
2	1	1
2	3	*NULL*

Fig. 1. Table $\langle\psi,R\rangle$

A
8

Fig. 2. Table $Sum_{A,R}(\langle\psi,R\rangle)$

B
6

Fig. 3. Table $Sum_{B,R}(\langle\psi,R\rangle)$

C
6

Fig. 4. Table $Sum_{C,R}(\langle\psi,R\rangle)$

Let $\leq$ be a linear order on the universal domain $\boldsymbol{D}$.

Definition 19. The aggregate operation $Min_{A,R}$ by the attribute A of the finite table on scheme R, $A\in R$, is a unary parametric operation of the form $Min_{A,R}:\Psi(R)\to\Psi(\{A\})$, $Min_{A,R}(\langle\psi,R\rangle)=\left\langle\left\{\left\{\langle A,Min(\alpha_A)\rangle\right\}^1\right\},\{A\}\right\rangle$, where $\langle\psi,R\rangle\in\Psi(R)$. The $Min(\alpha_A)$ function is applied to a column with attribute A in the table $\langle\psi,R\rangle$, the result obtained is the minimum value among values of α_A. In addition, $NULL$ values don't undertake in attention.

Thus, $Min:2_m^{\boldsymbol{D}}\to\boldsymbol{D}$,

$$Min(\alpha_A)=\begin{cases}NULL\ if\ \Theta(\alpha_A)=\varnothing;\\ NULL\ if\ \Theta(\alpha_A)=\{NULL\};\\ \min\{d\mid d\in\Theta(\alpha_A)\setminus\{NULL\}\}\ if\ \Theta(\alpha_A)\setminus\{NULL\}\neq\varnothing.\end{cases}$$

We have $Min(\varnothing_m)=NULL$, $Min(\{NULL^n\})=NULL$, $Min(\{d_1^{n_1},\dots,d_k^{n_k}\})=\min\{d_1,\dots,d_k\}$ if all elements d_i, $i=\overline{1,k}$, differ from $NULL$.

In the case of the empty table $\langle\psi_\varnothing,R\rangle$ we have $Min_{A,R}(\langle\psi_\varnothing,R\rangle)=\left\langle\left\{\left\{\langle A,NULL\rangle\right\}^1\right\},\{A\}\right\rangle$, here $\psi_\varnothing=\varnothing_m$.

Example 2. Let $\langle\psi,R\rangle$ be the table of Fig 1. Then $Min_{A,R}(\langle\psi,R\rangle)=\left\langle\left\{\left\{\langle A,2\rangle\right\}^1\right\},\{A\}\right\rangle$, $Min_{B,R}(\langle\psi,R\rangle)=\left\langle\left\{\left\{\langle B,0\rangle\right\}^1\right\},\{B\}\right\rangle$, $Min_{C,R}(\langle\psi,R\rangle)=\left\langle\left\{\left\{\langle C,1\rangle\right\}^1\right\},\{C\}\right\rangle$.

Definition 20. The aggregate operation $Max_{A,R}$ by the attribute A of the finite table on scheme R, $A\in R$, is a unary parametric operation of the form $Max_{A,R}:\Psi(R)\to\Psi(\{A\})$, $Max_{A,R}(\langle\psi,R\rangle)=\left\langle\left\{\left\{\langle A,Max(\alpha_A)\rangle\right\}^1\right\},\{A\}\right\rangle$, where $\langle\psi,R\rangle\in\Psi(R)$. The $Max(\alpha_A)$ function is applied to a column with attribute A in the table $\langle\psi,R\rangle$, the result obtained is the maximum value among values of α_A. In addition, $NULL$ values don't undertake in attention.

[b] The top index 1 specifies that the table include the tuple $\{\langle A,Sum(\alpha_A)\rangle\}$ only once, i.e. $\{\{\langle A,Sum(\alpha_A)\rangle\}^1\}$ is {1}-multiset.

doi:10.15849/icit.2015.0056 © ICIT 2015 (http://icit.zuj.edu.jo/ICIT15)

Thus, $Max : 2_m^{\boldsymbol{D}} \to \boldsymbol{D}$,

$$Max(\alpha_A) = \begin{cases} NULL \; if \;\; \Theta(\alpha_A) = \varnothing; \\ NULL \; if \;\; \Theta(\alpha_A) = \{NULL\}; \\ \max\{d \mid d \in \Theta(\alpha_A) \setminus \{NULL\}\} \; if \;\; \Theta(\alpha_A) \setminus \{NULL\} \neq \varnothing. \end{cases}$$

We have $Max(\varnothing_m) = NULL$, $Max(\{NULL^n\}) = NULL$, $Max\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right) = \max\{d_1, \ldots, d_k\}$ if all elements d_i, $i = \overline{1,k}$, differ from $NULL$.

In the case of the empty table $\langle \psi_\varnothing, R \rangle$ we have $Max_{A,R}\left(\langle \psi_\varnothing, R \rangle\right) = \left\langle \left\{\left\{\langle A, NULL \rangle\right\}^1\right\}, \{A\} \right\rangle$, here $\psi_\varnothing = \varnothing_m$.

Example 3. Let $\langle \psi, R \rangle$ be the table of Fig 1. Then

$Max_{A,R}\left(\langle \psi, R \rangle\right) = \left\langle \left\{\left\{\langle A, 2 \rangle^1\right\}\right\}, \{A\} \right\rangle$,

$Max_{B,R}\left(\langle \psi, R \rangle\right) = \left\langle \left\{\left\{\langle B, 3 \rangle\right\}^1\right\}, \{B\} \right\rangle$,

$Max_{C,R}\left(\langle \psi, R \rangle\right) = \left\langle \left\{\left\{\langle C, 3 \rangle\right\}^1\right\}, \{C\} \right\rangle$.

Definition 21. The aggregate operation $Count_{A,R}$ by the attribute A of the finite table on scheme R, $A \in R$, is a unary parametric operation of the form $Count_{A,R} : \Psi(R) \to \Psi(\{A\})$,

$Count_{A,R}\left(\langle \psi, R \rangle\right) = \left\langle \left\{\left\{\langle A, Count(\alpha_A) \rangle\right\}^1\right\}, \{A\} \right\rangle$, where $\langle \psi, R \rangle \in \Psi(R)$. The $Count(\alpha_A)$ function is applied to a column with attribute A in the table $\langle \psi, R \rangle$, the result obtained is the count of all values of α_A which differ from $NULL$.

Thus, $Count : 2_m^{\boldsymbol{D}} \to N$, $Count(\alpha_A) = \sum_{d \in \Theta(\alpha_A) \setminus \{NULL\}} \alpha_A(d)$. Put by definition that the sum of an empty set of elements is equal to zero.

So, we have $Count(\varnothing_m) = 0$, $Count(\{NULL^n\}) = 0$, $Count\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right) = n_1 + \ldots + n_k$ if all elements d_i, $i = \overline{1,k}$, differ from $NULL$.

In the case of the empty table $\langle \psi_\varnothing, R \rangle$ we have $Count_{A,R}\left(\langle \psi_\varnothing, R \rangle\right) = \left\langle \left\{\left\{\langle A, 0 \rangle\right\}^1\right\}, \{A\} \right\rangle$, here $\psi_\varnothing = \varnothing_m$.

Example 4. Let $\langle \psi, R \rangle$ be the table of Fig 1. Then

$Count_{A,R}\left(\langle \psi, R \rangle\right) = \left\langle \left\{\left\{\langle A, 4 \rangle\right\}^1\right\}, \{A\} \right\rangle$,

$Count_{B,R}\left(\langle \psi, R \rangle\right) = \left\langle \left\{\left\{\langle B, 5 \rangle\right\}^1\right\}, \{B\} \right\rangle$,

$Count_{C,R}\left(\langle \psi, R \rangle\right) = \left\langle \left\{\left\{\langle C, 4 \rangle\right\}^1\right\}, \{C\} \right\rangle$.

We assume that a numerical subset Num of the universal domain $\boldsymbol{D}$ is closed under the (partial operation) division operation $/ : Num \times Num \overset{\sim}{\to} Num$. We will determine the division operation so that when the first argument is equal to $NULL$ the function accepts value $NULL$.

Definition 22. The aggregate operation $Avg_{A,R}$ by the attribute A of the finite table on scheme R, $A \in R$, is a unary parametric operation of the form $Avg_{A,R} : \Psi(R) \to \Psi(\{A\})$,

$Avg_{A,R}\left(\langle \psi, R \rangle\right) = \left\langle \left\{\left\{\langle A, Avg(\alpha_A) \rangle\right\}^1\right\}, \{A\} \right\rangle$, where $\langle \psi, R \rangle \in \Psi(R)$. The $Avg(\alpha_A)$ function is applied to a column with attribute A in the table $\langle \psi, R \rangle$, the result obtained is the arithmetic mean of values in α_A which differ from $NULL$.

Thus, $Avg : 2_m^{Num} \to Num$ and $Avg(\alpha_A) = \dfrac{Sum(\alpha_A)}{Count(\alpha_A)}$.

We have $Avg(\varnothing_m) = \dfrac{Sum(\varnothing_m)}{Count(\varnothing_m)} = \dfrac{NULL}{0} = NULL$,

$Avg(\{NULL^n\}) = \dfrac{Sum(\{NULL^n\})}{Count(\{NULL^n\})} = \dfrac{NULL}{0} = NULL$,

$Avg(\{d_1^{n_1}, \ldots, d_k^{n_k}\}) = \dfrac{Sum(\{d_1^{n_1}, \ldots, d_k^{n_k}\})}{Count(\{d_1^{n_1}, \ldots, d_k^{n_k}\})} = \dfrac{\sum_{i=1}^{k} d_i n_i}{(n_1 + \ldots + n_k)}$ if all elements d_i, $i = \overline{1,k}$, differ from $NULL$.

In the case of the empty table $\langle \psi_\varnothing, R \rangle$ we have $Avg_{A,R}\left(\langle \psi_\varnothing, R \rangle\right) = \left\langle \left\{\left\{\langle A, NULL \rangle\right\}\right\}, \{A\} \right\rangle$, here $\psi_\varnothing = \varnothing_m$.

Example 5. Let $\langle \psi, R \rangle$ be the table of Fig 1. Then

$Avg_{A,R}\left(\langle \psi, R \rangle\right) = \left\langle \left\{\left\{\langle A, 2 \rangle\right\}^1\right\}, \{A\} \right\rangle$,

$$Avg_{B,R}\left(\langle \psi, R \rangle\right) = \left\langle \left\{\left\{\left\langle B, \frac{6}{5} \right\rangle\right\}^1\right\}, \{B\} \right\rangle,$$

$$Avg_{C,R}\left(\langle \psi, R \rangle\right) = \left\langle \left\{\left\{\left\langle C, \frac{6}{4} \right\rangle\right\}^1\right\}, \{C\} \right\rangle.$$

Definition 23. The aggregate operation $Count_{A,R}(*)$ by the attribute A of the finite table on scheme R, $A \in R$, is a unary parametric operation of the form $Count_{A,R}(*) : \Psi(R) \to \Psi(\{A\})$,

$Count_{A,R}(*)\left(\langle \psi, R \rangle\right) = \left\langle \left\{\left\{\langle A, \|\psi\| \rangle\right\}^1\right\}, \{A\} \right\rangle$, where $\langle \psi, R \rangle \in \Psi(R)$, and $\|\psi\|$ is the rank of the multiset ψ.

The operation $Count_{A,R}(*)$ finds the number of tuples in the table $\langle \psi, R \rangle$.

In the case of an empty table $\langle \psi_{\varnothing}, R \rangle$ we have $Count_{A,R}(*)(\langle \psi_{\varnothing}, R \rangle) = \left\langle \left\{ \left\{ \langle A, \|\varnothing_m\| \rangle \right\}^1 \right\}, \{A\} \right\rangle = \left\langle \left\{ \left\{ \langle A, 0 \rangle \right\}^1 \right\}, \{A\} \right\rangle$, here $\psi_{\varnothing} = \varnothing_m$.

Example 5. Let $\langle \psi, R \rangle$ be the table of Fig 1. Then

$$Count_{A,R}(*)(\langle \psi, R \rangle) = \left\langle \left\{ \left\{ \langle A, 5 \rangle \right\}^1 \right\}, \{A\} \right\rangle,$$

$$Count_{B,R}(*)(\langle \psi, R \rangle) = \left\langle \left\{ \left\{ \langle B, 5 \rangle \right\}^1 \right\}, \{B\} \right\rangle,$$

$$Count_{C,R}(*)(\langle \psi, R \rangle) = \left\langle \left\{ \left\{ \langle C, 5 \rangle \right\}^1 \right\}, \{C\} \right\rangle.$$

V. CONCLUSIONS

In this paper the multiset table algebra is considered. The signature of the multiset table algebra is filled up with new operations such as inner and outer join, semijoin and aggregate operations. The special element NULL is inserted in the universal domain for a define of outer operations.

It should also be noted that a parameter of aggregate operations is not necessarily only a single attribute; it also can be some function of the tuples.

[1] E. F. Codd, "A Relational Model for Large Shad Data Banks," in *Communications of the ACM*, USA, New York, 1970, vo1.13, No.6, pp. 377-387.

[2] Paul W.P.J. Grefen and Rolf A. de By, "A Multi-Set Extended Relational Algebra. A Formal Approach to a Practical Issue," in *10th International Conference on Data Engineering*, Houston, TX, USA, 1994, pp. 80-88.

[3] G. Lamperti, M. Melchiori, and M. Zanella, "On Multisets in Database Systems," in *Proceedings of the Workshop on Multiset Processing: Mathematical, Computer Science, and Molecular Computing Points of View*, London, UK, 2001, pp. 147-216.

[4] H. Garcia-Molina, J. D. Ullman, J. Widom, "Algebraic and Logical Query Languages," *Database Systems:The Complete Book*, in 2th ed. New Jersey, Upper Saddle River, 2009, ch. 2, sec. 5, pp. 203-241.

[5] V. Redko, J. Brona, D. Buy, S. Poliakov, "Compositional semantics of SQL," in *Relation Database: Relation Algebras and SQL-similar Languages*, Kyiv, Ukraine, 2001, ch. 3, sec. 3.4, pp. 151-180.

[6] D. B. Buy, I. M. Glushko, "Extended of Table Algebra: Multiset Table Algebra," in *Modern scientific research and their practical application*, Ukraine, Odessa, 2013, vol.J11309, Article CID Number 261.

[7] N. Cutland, "Prologue. Prerequisites and notation"in *Computability. An introduction to recursive function theory*. London, Cambridge University Press, 1980, sec. 2, pp. 2-4.