Proceedings of the of the 17th Conference on Computer Science and Intelligence Systems pp. 1–6
DOI: 10.15439/2022F300
ISSN 2300-5963 ACSIS, Vol. 30

Two new operations over extended index matrices and their applications in Big Data

Krassimir Atanassov*†, Veselina Bureva†
*Dept. of Bioinformatics and Mathematical Modelling
Institute of Biophysics and Biomedical Engineering,
Bulgarian Academy of Sciences
105 Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
†Intelligent Systems Laboratory
"Prof. Asen Zlatarov" University
Burgas–8010, Bulgaria
E-mail: krat@bas.bg, vbureva@btu.bg

***Abstract*—The Index Matrices (IMs) are extensions of the matrices of algebra. Over the IMs different operations, relations and operators are defined. In the present paper, two new operations over IMs are defined and some of their properties are studied. An application of these operations for describing of Big Data procedure is discussed. Hortonworks Data Platform (HDP) is used to provide capabilities for data warehouse processing in the Big Data environment. Apache Hive is selected for data warehouse construction and querying. Firstly, data warehouse for product sells is implemented. The tables for employees, customers, products and product sells are created. Thereafter the new index matrix operations for difference between two IMs are executed for the first time in the Big Data environment . SQL queries are written to demonstrate the operations. The new index matrix operations are executed using SQL JOIN notation and logical operator NOT EXISTS.**

***Index Terms*—Big Data, Extended index matrix, Operation.**

I. Introduction

THE CONCEPT of Index Matrix (IM) was introduced in 1984 and in more details - in 1987 [2], but the full description of the research over them was published in [3] exactly 30 years later.

Different extensions and modifications of the concept of an IM are described in [3]. One of them is an Extended IM (EIM), introduced firstly in [4]. They include as partial cases standard IM with elements of real or complex numbers, the $(0,1)$-IM with elements from set $\{0,1\}$, the logical IM with elements variables, propositions or predicates, the intuitionistic fuzzy IMs. The elements of the EIM can be each objects, in this number - whole IMs.

Different relations, operations and operators are defined over IMs and more general - over EIM. Only part of them have analogues in the theory of the standard matrices (see, e.g., [6], [7]).

Here, two new operations over EIMs are defined and some of their properties are studied. It will be obvious that these new operations can be transfer over each one of the partial cases of the EIM.

Firstly, we give the definition of an EIM.

Let $\mathcal{I}$ be a fixed set of indices,

$$\mathcal{I}^n = \{\langle i_1, i_2, ..., i_n\rangle | (\forall j : 1 \leq j \leq n)(i_j \in \mathcal{I})\}$$

and

$$\mathcal{I}^* = \cup_{1 \leq n \leq \infty} \mathcal{I}^n.$$

Let $\mathcal{X}$ be a fixed set of some objects. In the particular cases, they can be either real numbers, or only the numbers 0 or 1, or logical variables, propositions or predicates, etc.

Let operations $\circ, * : \mathcal{X} \times \mathcal{X} \to \mathcal{X}$ be fixed and let $(\mathcal{X}, \circ, e_\circ)$ and $(\mathcal{X}, *, e_*)$ be groups with unit elements $e_\circ$ and e_*, respectively. For example, when operation "$\circ$" is "+" or "-", $e_\circ$ will be 0, while when it is "$\times$" or ":" – it will be 1. In some cases, it is suitable to define the unit element by $\perp$ and it to be an empty object.

An EIM with index sets K and L $(K, L \subset \mathcal{I}^*)$ and elements from set $\mathcal{X}$ is called the object (see, [4], [3]):

$$[K, L, \{a_{k_i,l_j}\}] \equiv$$

	l_1	...	l_j	...	l_n
k_1	a_{k_1,l_1}	...	a_{k_1,l_j}	...	a_{k_1,l_n}
⋮	⋮	...	⋮	...	⋮
k_i	a_{k_i,l_1}	...	a_{k_i,l_j}	...	a_{k_i,l_n}
⋮	⋮	...	⋮	...	⋮
k_m	a_{k_m,l_1}	...	a_{k_m,l_j}	...	a_{k_m,l_n}

,

where $K = \{k_1, k_2, ..., k_m\}$, $L = \{l_1, l_2, ..., l_n\}$, for $1 \leq i \leq m$, and $1 \leq j \leq n : a_{k_i,l_j} \in \mathcal{X}$.

In [3], for the IMs $A = [K, L, \{a_{k_i,l_j}\}], B = [P, Q, \{b_{p_r,q_s}\}]$, operations that are analogous to the usual matrix operations of addition and multiplication are defined, as well as other, specific ones. Here, we give only three of these operations, that will be used below.

Addition

$$A \oplus_{(\circ)} B = [K \cup P, L \cup Q, \{c_{t_u,v_w}\}],$$

where

$$c_{t_u,v_w} = \begin{cases} a_{k_i,l_j}, & \text{if } t_u = k_i \in K \text{ and } v_w = l_j \in L - Q \\ & \text{or } t_u = k_i \in K - P \text{ and } v_w = l_j \in L; \\ b_{p_r,q_s}, & \text{if } t_u = p_r \in P \text{ and } v_w = q_s \in Q - L \\ & \text{or } t_u = p_r \in P - K \text{ and } v_w = q_s \in Q; \\ a_{k_i,l_j} \circ b_{p_r,q_s}, & \text{if } t_u = k_i = p_r \in K \cap P \\ & \text{and } v_w = l_j = q_s \in L \cap Q \\ 0, & \text{otherwise} \end{cases}$$

Structural subtraction

$$A \ominus B = [K - P, L - Q, \{c_{t_u,v_w}\}],$$

where here and below "–" is the set–theoretic difference operation and

$$c_{t_u,v_w} = a_{k_i,l_j}, \text{ for } t_u = k_i \in K - P \text{ and } v_w = l_j \in L - Q.$$

Projection

$$pr_{M,N} A = [M, N, \{b_{k_i,l_j}\}],$$

where $M \subseteq K$, $N \subseteq L$, and for each $k_i \in M$ and each $l_j \in N$, $b_{k_i,l_j} = a_{k_i,l_j}$.

II. Definitions of the two new operations

Let the two EIMs $A = [K, L, \{a_{k_i,l_j}\}]$ and $B = [P, Q, \{b_{p_r,q_s}\}]$ are given.

The first, simpler, operation is defined by

$$A \overset{\bullet}{_1} B = [K \div P, L \div Q, \{c_{u_v,w_t}\}],$$

where and below for every two arbitrary sets X, Y:

$$X \div Y = (X - Y) \cup (Y - X);$$

$$c_{u_v,w_t} = \begin{cases} a_{k_i,l_j}, & \text{if } u_v = k_i \in K - P \text{ and } w_t = l_j \in L - Q \\ b_{p_r,q_s}, & \text{if } w_t = p_r \in P - K \text{ and } w_t = q_s \in Q - L \\ \perp & \text{otherwise} \end{cases}$$

The second operation is defined by

$$A \overset{\bullet}{_2} B = [K \cup P, L \cup Q, \{c_{u_v,w_t}\}],$$

where

$$c_{u_v,w_t} = \begin{cases} a_{k_i,l_j}, & \text{if } u_v = k_i \in K - P \text{ and } w_t = l_j \in L \\ & \text{or } u_v = k_i \in K \text{ and } w_t = l_j \in L - Q \\ b_{p_r,q_s}, & \text{if } w_t = p_r \in P - K \text{ and } w_t = q_s \in Q \\ & \text{if } w_t = p_r \in P \text{ and } w_t = q_s \in Q - L \\ \perp & \text{otherwise} \end{cases}$$

The geometrical interpretations of both operations are shown on Figures 1 and 2.

From the definitions and geometrical interpretations of both operations we see immediately that the following assertions are valid.

Theorem 1. For every two EIMs A and B, for each operation $\circ$, and for operation $*$ defined for every two $x, y \in \mathcal{X}$ by $x * y = e_*$ there are follows:

$$A \overset{\bullet}{_1} B = (A \ominus B) \oplus_{(\circ)} (B \ominus A),$$

$$A \overset{\bullet}{_2} B = A \oplus_{(*)} B.$$

Proof. For the definitions of the operations $\ominus$ and $\oplus_\circ$ we obtain

$$(A \ominus B) \oplus_{(\circ)} (B \ominus A) = [K - P, L - Q, \{c_{t_u,v_w}\}] \oplus_{(\circ)} [P - K, Q - L, \{d_{e_f,g_h}\}$$
$$= [(K - P) \cup (P - K), (L - Q) \cup (Q - L), \{c'_{t'_{u'},v'_{w'}}\}]$$
$$= [K \div P, L \div Q, \{c'_{t'_{u'},v'_{w'}}\}] = A \div B,$$

because

$$c'_{t'_{u'},v'_{w'}} = \begin{cases} c_{t_u,v_w} = a_{k_i,l_j}, \text{ for } t'_{u'} = t_u = k_i \in K - P \text{ and } v'_{w'} = v_w \\ d_{e_f,g_h} = b_{p_r,q_s}, \text{ for } t'_{u'} = e_f = p_r \in P - K \text{ and } v'_{w'} = g_h \end{cases}$$

Obviously, operation $\circ$ cannot be applied over the elements of both EIMs because there are not at least two elements from the both EIMs that have equal indices.

The second equality is proved by a similar way. □

Theorem 2. For every two EIMs A and B, for each operation $\circ$:

$$A \overset{\bullet}{_1} B = pr_{K-P,L-Q} A \oplus_{(\circ)} pr_{P-K,Q-L} B,$$

$$A \overset{\bullet}{_2} B = pr_{K-P,L} A \oplus_{(\circ)} pr_{K \cap P, L \div Q} A \oplus_{(\circ)} pr_{P-K,Q} A$$
$$= pr_{K,L-Q} A \oplus_{(\circ)} pr_{K \div P, L \cap Q} A \oplus_{(\circ)} pr_{P,Q-L} A.$$

The proof is similar to the above one.

Let $\mathcal{M}$ be the set of all EIMs with elements from $\mathcal{X}$[1].

Let

$$I_\emptyset = [\emptyset, \emptyset, \{\perp\}].$$

Theorem 3. $(\mathcal{M}, \overset{\bullet}{_1}, I_\emptyset)$ is a commutative group.

Proof. From the definition of operation "$\overset{\bullet}{_1}$" it follows directly that for each two EIMs $A, B \in \mathcal{M}$, $A \overset{\bullet}{_1} B \in \mathcal{M}$.

Using the well-known equality for every three sets X, Y and Z:

$$(X \div Y) \div Z = X \div (Y \div Z),$$

for every three EIMs $A, B, C \in \mathcal{M}$, where

$$C = [D, E, \{c_{d_f,e_g}\}]$$

we obtain

$$(A \overset{\bullet}{_1} B) \overset{\bullet}{_1} C = ([K, L, \{a_{k_i,l_j}\}] \overset{\bullet}{_1} [P, Q, \{b_{p_r,q_s}\}]) \overset{\bullet}{_1} [D, E, \{c_{d_f,e_g}\}]$$
$$= [K \div P, L \div Q, \{x_{u_v,w_t}\}] \overset{\bullet}{_1} [D, E, \{c_{d_f,e_g}\}]$$

[1]When $\mathcal{X}$ is a set (class) in the sense of NBG-set theory of all predicates, then $\mathcal{M}$ will be a set (class).

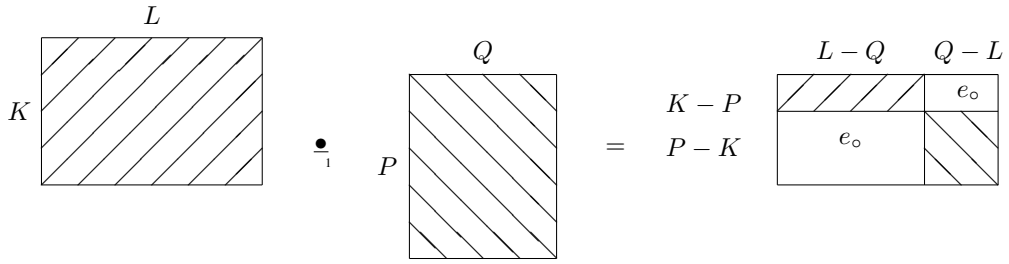

Fig. 1. Geometrical interpretation of operation $\overset{\bullet}{\div}_1$

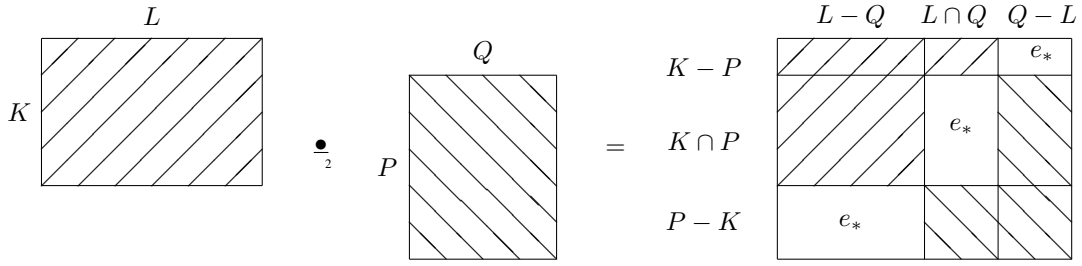

Fig. 2. Geometrical interpretation of operation $\overset{\bullet}{\div}_2$

(where each element x_{u_v,w_t} is some element a_{k_i,l_j} or some element b_{p_r,q_s})

$$= [(K \div P) \div D, (L \div Q) \div E, \{y_{\alpha_\beta,\gamma_\delta}\}]$$

(where each element $y_{\alpha_\beta,\gamma_\delta}$ is some element a_{k_i,l_j} or some element b_{p_r,q_s}), or some element c_{d_f,e_g})

$$= [K \div (P \div D), L \div (Q \div E), \{y_{\alpha_\beta,\gamma_\delta}\}]$$

$$= ([K, L, \{a_{k_i,l_j}\}] \overset{\bullet}{\div}_1 ([P \div D, Q \div E, \{z_{\varepsilon_\zeta,\eta_\theta}\}])$$

(where each element $z_{\varepsilon_\zeta,\eta_\theta}$ is some element b_{p_r,q_s}), or some element c_{d_f,e_g})

$$= [K, L, \{a_{k_i,l_j}\}] \overset{\bullet}{\div}_1 ([P, Q, \{b_{p_r,q_s}\}]) \overset{\bullet}{\div}_1 [D, E, \{c_{d_f,e_g}\}]).$$

Now, for EIM $I_\emptyset$ we obtain

$$A \overset{\bullet}{\div}_1 I_\emptyset = [K \div \emptyset, L \div \emptyset, \{c_{u_v,w_t}\}]$$

(where each element c_{u_v,w_t} coincides with the element a_{k_i,l_j} from A for $u_v = k_i, w_t = l_j$)

$$= [K, L, \{a_{k_i,l_j}\}] = A.$$

Analogously is checked that

$$I_\emptyset \overset{\bullet}{\div}_1 A = A.$$

From the well-known equality $X \div Y = Y \div X$ it follows that

$$A \overset{\bullet}{\div}_1 B = [K \div P, L \div Q, \{c_{u_v,w_t}\}]$$

$$m = [P \div K, Q \div L, \{c_{u_v,w_t}\}] = B \overset{\bullet}{\div}_1 A,$$

i.e., the operation $\overset{\bullet}{\div}_1$ is commutative.

Finally,

$$A \overset{\bullet}{\div}_1 A = [K \div K, L \div L, \{x_{u_v,w_t}\}] = [\emptyset, \emptyset, \{x_{u_v,w_t}\}] = I_\emptyset,$$

because of lack of indices, element x_{u_v,w_t} must be $\perp$. □

Theorem 4. $(\mathcal{M}, \overset{\bullet}{\div}_2, I)$ is a commutative monoid.

The proof is similar to this of Theorem 3, without the last part, i.e.,as above, we can check that $(\mathcal{M}, \overset{\bullet}{\div}_2, I)$ is a commutative monoid, but it is not a group because the fact that there is not a set Y that for some non-empty set X to be valid $X \cup Y = \emptyset$.

III. An example of difference operationsin the Big Data environment

Hortonworks Data Platform (HDP) is a an open source framework for distributed storage and processing of huge datasets retrieving from different sources. HDP is used to discover insights from structured and unstructured data in the cloud or on-premises.It includes Big Data tools as Hadoop,

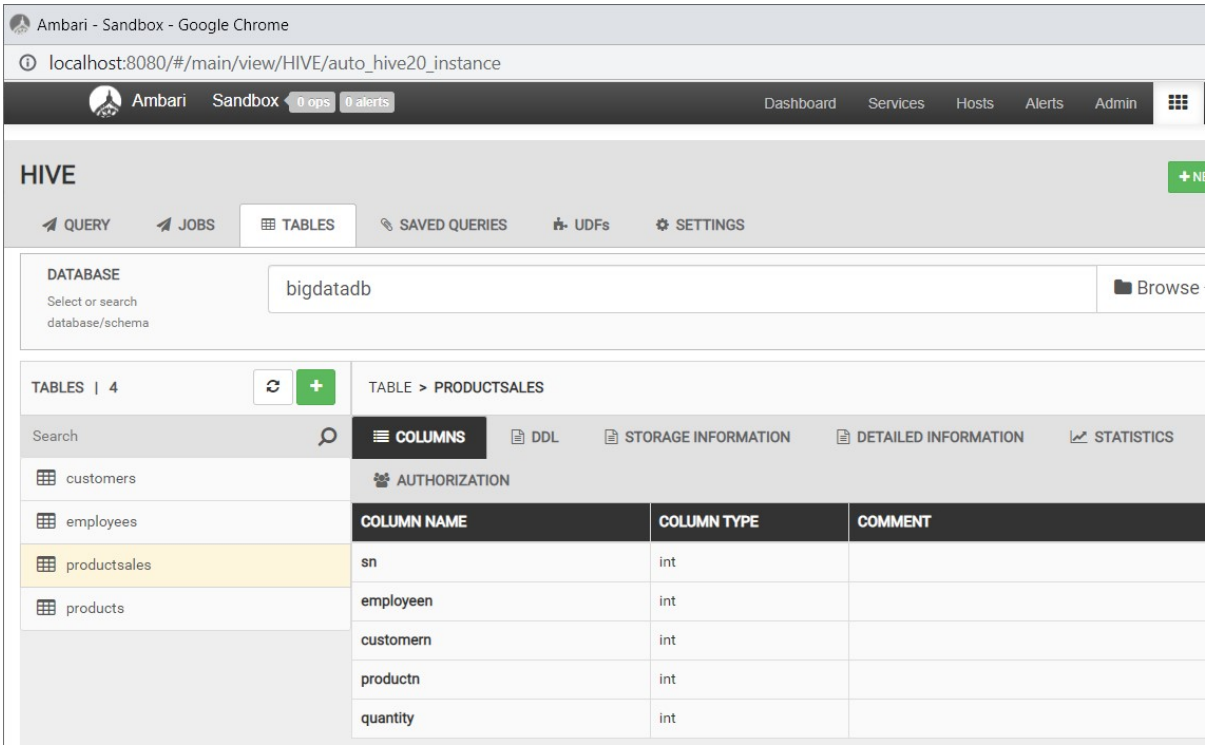

Fig. 3. Data warehouse bigdatadb view in Apache Hive environment

Yarn, MapReduce, Hbase, Hive, Flume, Kafka, Druid [5]. The example of the difference operations is performed in the environment of Apache Hive. Apache Hive is a data warehouse used for reading, writing, and managing large amounts of data stored in distributed storage in SQL. Apache Hive flexibility can be extended using the used-defined functions (UDF) [1]. In the current investigation a data warehouse for product sells is implemented. The tables are uploaded using previously prepared csv files. Apache Hive supports only the sets operations $Union\ All$ and $Union\ [Distinct]$. The $Intersect$ and $Except\ (Minus)$ operations are not included. In the current investigation for the first time the new index matrix operations performing operation difference will be applied by the analogy of the SQL $JOIN$ clause and the logical operator $NOT\ EXISTS$ in the Big Data environment. The tables $Customers$ and $Employees$ from *bigdatadb* data warehouse (Fig. 3). The authors will not compare the indexed field Number – the comparison will be performed using the columns for the first name and the last name of the tables. There are from the same data type. The tables $Employees$ (Fig. 4) and $Customers$ (Fig. 5) have the following from:

An example of the standard JOIN operation has the following form:

```
SELECT FName, Lname, CFname, CLName
FROM Employees JOIN ProductSales
ON Employees.Number=ProductSales.EmployeeN
JOIN Customers
ON Customers.CNumber=ProductSales.CustomerN
```

employees.number	employees.fname	employees.lname	employees.town
1	Ivan	Dimitrov	Burgas
2	George	Radev	Burgas
3	Simona	Ivanova	Sofia
4	Radi	Hristov	Burgas
5	Vasilena	Moneva	Sofia
6	Hristo	Rachev	Burgas
7	Qni	Simov	Burgas
8	Radina	Tomova	Sofia
9	Stanislav	Stoyanov	Sofia
10	Radostina	Dimitrova	Burgas
11	Dimo	Dimov	Burgas
12	Maria	Yaneva	Burgas
13	Dimitar	Stoilov	Sofia
14	Svetla	Iliyanova	Burgas
15	Margarita	Simova	Burgas
16	Radomira	Kirova	Sofia
17	Radovesta	Todorova	Sofia
18	Valeria	Taneva	Sofia
19	Silviya	Stancheva	Burgas
20	Stilyan	Stoilov	Burgas

Fig. 4. Table of Employees

customers.cnumber	customers.cfname	customers.clname	customers.ctown
1	Ivan	Dimitrov	Burgas
2	George	Radev	Burgas
3	Siyana	Tumova	Sofia
4	Radina	Mihaylova	Sofia
5	Stancho	Dimitrov	Burgas
6	Hari	Dimov	Burgas
7	Qni	Simov	Burgas
8	Radina	Tomova	Sofia
9	Mihail	Radev	Burgas
10	Hrisi	Staneva	Sofia
11	Dimo	Dimov	Burgas
12	Maria	Yaneva	Burgas
13	Monika	Ganeva	Burgas
14	Stanimit	Ganeva	Sofia
15	Hristo	ivanov	Sofia
16	Iliyan	Fotev	Sofia
17	Hristiana	Racheva	Burgas
18	Valeria	Taneva	Sofia
19	Silviya	Stancheva	Burgas
20	Stilyan	Stoilov	Burgas

Fig. 5. Table of Customers

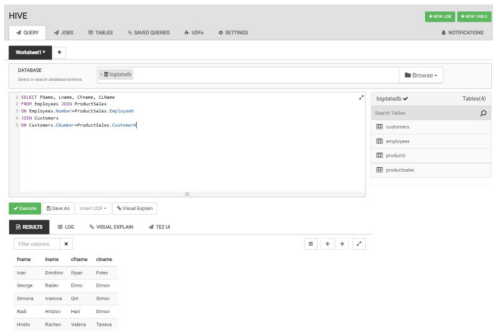

Fig. 6. Execution of standard JOIN operation

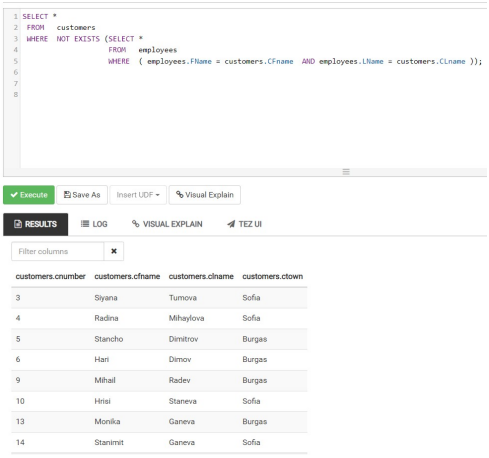

Fig. 7. SQL query for difference operation – case 1

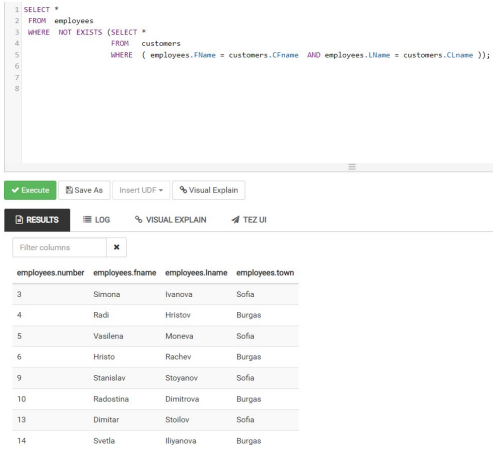

Fig. 8. SQL query for difference operation – case 2

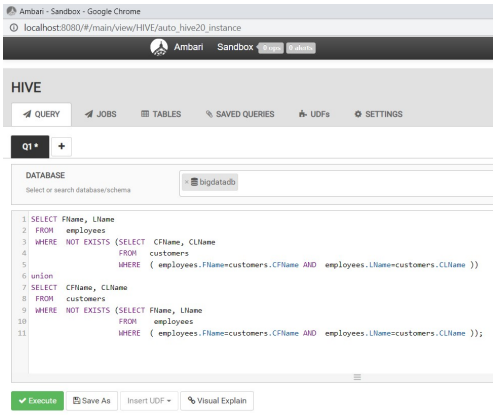

Fig. 9. SQL query for difference operation – case 3

The result presents the names of the customers and the employees. The SQL query in the Apache Hive environment is presented on the Fig. 6.

The simulation of difference operation is performed using the $NOT\ EXIST$ logical operator in the $SELECT$ statement and comparison of the desired columns from the tables in the sub-query (Fig. 7).

```
SELECT * FROM customers
WHERE  NOT EXISTS (SELECT *
FROM employees
WHERE  (employees.FName = customers.CFname
    AND employees.LName = customers.CLname)
       );
```

The result of the query presents customers not included in the list of the employees. The columns for first names and last names are compared.

The same query with replaced tables in the query and sub-query can be executed to receive the employees not included in the table of the customers (Fig. 8).

The result of the query is presented on the Fig. 10. It combines the employees that are not customers and the customers that are not employees.

The combination of the previous queries using the UNION operator is executed (Fig. 9).

Fig. 10. Result of SQL query for difference operation – case 3

```
SELECTF Name, LName
FROM employees
WHERE NOT EXISTS (SELECT CFName, CLName
FROM customers
WHERE  (employees.FName=customers.CFName
   AND  employees.LName=customers.CLName )
   )
UNION
SELECT CFName, CLName
FROM customers
WHERE NOT EXISTS (SELECT FName, LName
FROM employees
WHERE  (employees.FName=customers.CFName
   AND  employees.LName=customers.CLName )
);
```

IV. Conclusion

Two new operations are introduced in the paper. In future, other properties of them will be studied and some other applications will be discussed. A part of them will be related to Big Data and Data Mining as continuation of the discussed in the present paper.

Acknowledgment

The authors acknowledge the support from the project UNITe BG05M2OP001-1.001-0004 /28. 02.2018 (2018-2023).

References

[1] Apache Hive, https://hive.apache.org/
[2] Atanassov K. Generalized index matrices, Comptes rendus de l'Academie Bulgare des Sciences, Vol. 40, 1987, No. 11, 15–18.
[3] Atanassov, K., Index Matrices: Towards an Augmented Matrix Calculus, Springer, Cham, 2014.
[4] Atanassov, K., Extended index matrices, Proc. of the 7-th IEEE Conference "Intelligent Systems", Warsaw, 24-26 Sept. 2014 (P. Angelov, K. Atanassov, L. Doukovska, M. Hadjiski, V. Jotsov, J. Kacprzyk, N. Kasabov, S. Sotirov, E. Szmidt, S. Zadrozny, Eds.), Springer, Cham, 2015, 305-317.
[5] Hortonworks Data platform, https://www.cloudera.com/products/hdp.html
[6] Lankaster, P. Theory of Matrices, Academic Press, New York, 1969.
[7] Zhang, F. Matrix Theory. Springer, New York, 2011.