Доклади на Българската академия на науките
Comptes rendus de l'Académie bulgare des Sciences
Tome 75, No 5, 2022

ENGINEERING SCIENCES
Artificial intelligence

INDEX MATRIX REPRESENTATION OF BIG DATA STRUCTURES

Krassimir Atanassov[*,**], **Veselina Bureva**[**]

Received on November 25, 2021
Accepted on November 30, 2021

Abstract

It is shown that the apparatus of the Index Matrices (IM) can be used for representation of a BigData (BD) structure. An algorithm for searching of facts in the IM-form of a given BD is described. Ideas for further extensions of the BD-structure and of its facts, are formulated.

Key words: Big Data, fact, Index matrix
2020 Mathematics Subject Classification: 03E72, 11C20, 62R07

1. Introduction – short remarks on Big Data. Big data systems aim to deal with the opportunity of high velocity, high volumes, and high varieties of datasets. The data sources in big data system are presented in three different formats: structured data, semi-structured data and unstructured data. In the beginning the big data framework Apache Hadoop is developed [1]. The opportunity of storing different types of datasets is solved by NoSQL databases [2]. Thereafter the NewSQL databases are introduced to provide the SQL opportunities in big data environment [3]. In the end in-memory databases appear to provide faster and more predictable performance [4]. The big data systems and their applications are investigated in [5–10]. Big data analytics are performed using different query engines and big data platforms [11–13]. The data science methods are discussed in [14].

The authors acknowledge the support from the project UNITe BG05M2OP001-1.001-0004/28.02.2018 (2018–2023).
DOI:10.7546/CRABS.2022.05.12

The aim of the present paper is to describe a representation of a BD structure, using the apparatus of the Index Matrices (IMs) and to prove that this representation uses essentially smaller volume of the computer memory. An algorithm for searching of facts in a BD structure represented by an IM, is given.

2. Short remark on index matrices. The concept of an Index Matrix (IM) was introduced in [15] and described in details in [16].

Let $\mathcal{I}$ be a fixed set of indices and let $\mathcal{X}$ be a fixed set of some objects.

We call the object $[K, L, \{a_{k_i,l_j}\}]$ with index sets K and L ($K, L \subset \mathcal{I}$), an IM. It has the form

$$[K, L, \{a_{k_i,l_j}\}] \equiv \begin{array}{c|cccccc} & l_1 & \dots & l_j & \dots & l_n \\ \hline k_1 & a_{k_1,l_1} & \dots & a_{k_1,l_j} & \dots & a_{k_1,l_n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ k_i & a_{k_i,l_1} & \dots & a_{k_i,l_j} & \dots & a_{k_i,l_n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ k_m & a_{k_m,l_1} & \dots & a_{k_m,l_j} & \dots & a_{k_m,l_n} \end{array},$$

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

When $\mathcal{X}$ is a set of real numbers; or only the numbers 0 or 1; or logical variables, propositions or predicates, then it is a standard IM, while when $\mathcal{X}$ is a set of arbitrary objects, and in particular case – entire IMs, then it is called an extended IM (see [16]). Over IMs different operations, extending these over the standard matrices, as well as relations and operators are defined in [16]. There, 3-dimensional IMs are described, while in [17, 18] n-dimensional IMs are introduced.

In practice, below, we will use a 2-dimensional extended IM, but for brevity will denote it only as an IM.

3. On an IM representation of a BD. Let us have a BD with objects (facts) having the form

$$\left\langle F(\text{name}), \kappa_1^{F(\text{name})}, \kappa_2^{F(\text{name})}, \kappa_{s(F(\text{name}))}^{F(\text{name})}, F(\text{contents}) \right\rangle,$$

where $\kappa_1^{F(\text{name})}, \kappa_2^{F(\text{name})}, \dots, \kappa_{s(F(\text{name}))}^{F(\text{name})} \in \mathcal{K}$ is the set of all keywords that the facts of the BD can have at the present moment.

Let the set $\mathcal{K}$ be clustered in $k > 1$ clusters of keywords $K_1, K_2, \dots, K_k$.

Let the BD-facts be collected in n_1 clusters $C_1^1, C_2^1, \dots, C_{n_1}^1$, each one of which, let it be $C_{i_1}^1$ for some i_1: $1 \leq i_1 \leq n_1$, can be on a final level, i.e., it contains only facts, or it contains sub-clusters $C_1^{2,i_1}, C_2^{2,i_1}, \dots, C_{n_{i_1}}^{2,i_1}$ and a set of facts $S_{i_1}^{1,j}$ and each one of these facts contains keywords from the cluster K_j.

Let the BD have clusters on $q \geq 1$ levels. Therefore, we can construct the IM

$$[C_1, \mathcal{K}, \{a_{k_i,l_j}\}] \equiv \begin{array}{c|cccccc} & K_1 & \dots & K_j & \dots & K_k \\ \hline C_1^1 & a_{C_1^1,K_1} & \dots & a_{C_1^1,K_j} & \dots & a_{C_1^1,K_k} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ C_{i_1}^1 & a_{C_{i_1}^1,K_1} & \dots & a_{C_{i_1}^1,K_j} & \dots & a_{C_{i_1}^1,K_k} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ C_{n_1}^1 & a_{C_{n_1}^1,K_1} & \dots & a_{C_{n_1}^1,K_j} & \dots & a_{C_{n_1}^1,K_k} \end{array},$$

where for $1 \leq i_1 \leq n_1$ and for $1 \leq j \leq k$:

$$a_{C_{i_1}^1,K_j} = \begin{cases} S_{i_1}^{1,j}, \{C_1^{2,i_1}, C_2^{2,i_1}, \dots, C_{n_{2,i_1}}^{2,i_1}\}, & \text{if the cluster } C_{i_1}^1 \text{ is not from a final level and in each of its sub-clusters } C_1^{2,i_1}, C_2^{2,i_1}, \dots, C_{n_{2,i_1}}^{2,i_1} \text{ there is at least one fact that has keywords from the cluster } K_j, \\ S_{i_1}^{1,j}, & \text{if the cluster } C_{i_1}^1 \text{ is from a final level and the set } S_{i_1}^{1,j} \text{ contains all facts that have keywords from the cluster } K_j, \\ \varnothing & \text{if the cluster } C_{i_1}^1 \text{ is from a final level and in none of its sub-clusters there is a fact having keywords from the cluster } K_j. \end{cases}$$

In the same way, when we have a cluster $C_{i_p}^{p,i_1,\dots,i_{p-1}}$ from p-th level, where $1 \leq p \leq q-1$, we can determine its sub-clusters of $(p+1)$-st level (see Fig. 1) and we can construct the IM

$$[C_{i_p}^{p,i_1,\dots,i_{p-1}}, \{K_1, K_2, \dots, K_k\}, \{a_{k_i,l_j}\}]$$

$$\equiv \begin{array}{c|cccccc} & K_1 & \dots & K_j & \dots & K_k \\ \hline C_1^{p+1,i_1,\dots,i_p} & a_{C_1^1,K_1} & \dots & a_{C_1^1,K_j} & \dots & a_{C_1^1,K_k} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ C_{i_1}^1 & a_{C_{i_1}^1,K_1} & \dots & a_{C_{i_1}^1,K_j} & \dots & a_{C_{i_1}^1,K_k} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ C_{n_{p+1,i_1,\dots,i_p}}^{p+1,i_1,\dots,i_p} & a_{C_{n_1}^1,K_1} & \dots & a_{C_{n_1}^1,K_j} & \dots & a_{C_{n_1}^1,K_k} \end{array}.$$

Now, we will describe the BD-procedure related to searching of a fact from the BD.

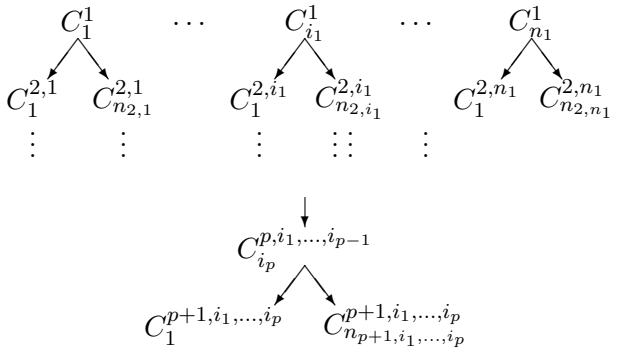

Fig. 1

Let the fact $\left\langle F(\text{name}), \kappa_1^{F(\text{name})}, \kappa_2^{F(\text{name})}, \kappa_{s(F(\text{name}))}^{F(\text{name})}, F(\text{contents}) \right\rangle$ have keywords from clusters $K_{j_1}, K_{j_2}, \ldots, K_{j_t}$, where $\{K_{j_1}, K_{j_2}, \ldots, K_{j_t}\} \subset \{K_1, K_2, \ldots, K_k\}$. Obviously, $t < k$.

The algorithm is the following.

Step 1. For each $u\colon 1 \le u \le t$ and for each cluster $C_{i_1}^1$ $(1 \le i_1 \le n_1)$ from first level, we determine all sets of clusters from second level $\{C_{v_1}^{2,i_1}, C_{v_2}^{2,i_1}, \ldots, C_{v_w}^{2,i_1}\}$, where $\{v_1, v_2, \ldots, v_w\} \subseteq \{1, 2, \ldots, n_{2,i_1}\}$ so that each one of these clusters contains a fact with a keyword from cluster K_{j_u}. Of course, it is possible some of these $C_{i_1}^1$-clusters to be empty.

Step 2. We construct the set

$$\bigcup_{i_1=1}^{n_1} \{C_{v_1}^{2,i_1}, C_{v_2}^{2,i_1}, \ldots, C_{v_w}^{2,i_1}\}$$

of all clusters from the second level in each one of which there exists at least one fact containing keyword from cluster K_{j_u}.

Step 3. We construct the set

$$C_2 = \bigcap_{p=1}^{t} \bigcup_{i_1=1}^{n_1} \{C_{v_1}^{2,i_1}, C_{v_2}^{2,i_1}, \ldots, C_{v_w}^{2,i_1}\} = \{C_1^2, C_2^2, \ldots, C_{m_2}^2\}.$$

Step 4. We construct the set

$$S_{1,j_1,j_2,\ldots,j_t} = \bigcap_{p=1}^{t} \bigcup_{i_1=1}^{n_1} S_{i_1}^{1,j_p}.$$

Step 5. We construct the IM

	K_{j_1}	$\dots$	K_{j_p}	$\dots$	K_{j_t}
C_1^2	$a_{C_1^2,K_{j_1}}$	$\dots$	$a_{C_1^2,K_{j_p}}$	$\dots$	$a_{C_1^2,K_{j_t}}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$C_{i_2}^2$	$a_{C_{i_2}^2,K_{j_1}}$	$\dots$	$a_{C_{i_2}^2,K_{j_p}}$	$\dots$	$a_{C_{i_2}^2,K_{j_t}}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$C_{n_2}^2$	$a_{C_{n_2}^2,K_{j_1}}$	$\dots$	$a_{C_{n_2}^2,K_{j_p}}$	$\dots$	$a_{C_{n_2}^2,K_{j_t}}$

.

We repeat the procedure, until we have measured up the last (q-th) level, for which we construct the IM

$[C_{i_q}^{q,i_1,\dots,i_{q-1}}, \{K_{j_1}, K_{j_2}, \dots, K_{j_t}\}, \{a_{k_i,l_j}\}]$

$$= \begin{array}{c|ccccc} & K_{j_1} & \dots & K_{j_p} & \dots & K_{j_t} \\ \hline C_1^{q,i_1,\dots,i_{q-1}} & a_{C_1^{q,i_1,\dots,i_{q-1}},K_{j_1}} & \dots & a_{C_1^{q,i_1,\dots,i_{q-1}},K_{j_p}} & \dots & a_{C_1^{q,i_1,\dots,i_{q-1}},K_{j_t}} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ C_{i_q}^{q,i_1,\dots,i_{q-1}} & a_{C_{i_q}^{q,i_1,\dots,i_{q-1}},K_{j_1}} & \dots & a_{C_{i_q}^{q,i_1,\dots,i_{q-1}},K_{j_p}} & \dots & a_{C_{i_q}^{q,i_1,\dots,i_{q-1}},K_{j_t}} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ C_{n_q}^{q,i_1,\dots,i_{q-1}} & a_{C_{n_q}^{q,i_1,\dots,i_{q-1}},K_{j_1}} & \dots & a_{C_{n_q}^{q,i_1,\dots,i_{q-1}},K_{j_p}} & \dots & a_{C_{n_q}^{q,i_1,\dots,i_{q-1}},K_{j_t}} \end{array},$$

where for $1 \leq i_q \leq n_q$ and for $1 \leq p \leq t$:

$$a_{C_{i_q}^{q,i_1,\dots,i_{q-1}},K_{j_p}} = \begin{cases} S_{i_q}^{q,i_1,\dots,i_{q-1}}, & \text{if in the cluster } C_{i_q}^{q,i_1,\dots,i_{q-1}} \text{ there is a fact} \\ & \text{having a keyword from the cluster } K_{j_p}, \\ \varnothing & \text{otherwise,} \end{cases}$$

where $S_{i_q}^{q,i_1,\dots,i_{q-1}}$ is the set of facts that contain a keyword from the cluster K_j. Now, we construct the set S that is the union of all S-sets. It contains all facts of the BD that have keywords from the clusters $K_{j_1}, K_{j_2}, \dots, K_{j_t}$.

This procedure is the basic one for all other BD-procedures. For example, it is used if the task is to change existing fact with keywords from the clusters $K_{j_1}, K_{j_2}, \dots, K_{j_t}$ with a new one, having the same keywords. Also, it will be used if the task is to erase the existing in the BD fact with keywords from the clusters $K_{j_1}, K_{j_2}, \dots, K_{j_t}$.

The most interesting case is when we must put a new fact, having keywords from the clusters $K_{j_1}, K_{j_2}, \dots, K_{j_t}$. Now, there are two cases.

In the first case, for the new fact there is no condition in which cluster it has to stay. Then, we extend all S-sets of the BD-clusters, having the same keywords, as the new fact.

In the second case, when for the new fact there is information in which cluster of the BD it must be put, then we extend only this S-set that is element of the indicated cluster.

The most interesting case is when the new fact contains keywords that do not exists in the set $\mathcal{K}$. Let these keywords be elements of a new cluster L. Then we construct a new IM $[C_1, \mathcal{K} \cup L, \{a_{k_i,l_j}\}]$ and work with it.

Finally, we discuss the size of the BD organized as a standard matrix M and as an IM.

Let f be the number of all BD-facts and let the number of p-th cluster K_p be k_p. Therefore, the number of all keywords is $h = \sum_{p=1}^{k} k_p$. Then, in the standard case the size of M is fh. In the IM-approach, the number of the IM-elements will be no more than fk, because each fact will be met in some row of the IMs, representing the respective cluster of facts, but, as seen from the above algorithm, we work only with the keyword-clusters that are k in number. Obviously,

$$k \ll h = \sum_{p=1}^{k} k_p.$$

4. Conclusion: ideas for the future. As it was shown, the volume of the BD-structure represented by IMs is smaller than the volume of a standard BD-structure. Program realization of the IM operations, relations and operators, and the proposed algorithm, is being prepared at the moment. In the future it will be incorporated in some BD-structures.

In [19] an expert system was described whose facts have the form:

$$\langle F, t_1, t_2, \ldots, t_n \rangle,$$

where the t-components correspond to time-moments: t_{2k-1} is the time-moment, in which the fact starts to be valid and t_{2k} is the time-moment, in which the fact stops to be valid. So, the expert system can answer different questions, e.g.: *"Is the fact valid now?", "Has the fact been valid once?", "Has the fact been valid in the past?", "Has the fact been valid sometimes?", "Has the fact been always valid?", "Has the fact been valid often?", "Has the fact been rarely valid?"*, etc. Now, this idea can be transformed in the case of BD and it can obtain IM-representation, too.

In the present research, we used a 2-dimensional IM that corresponds to the standard matrices. In [16], 3-dimensional and in [17] n-dimensional IMs are described. So, by them, we can describe n-dimensional BD in the same manner, as above.

REFERENCES

[1] White T. (2015) Hadoop: The Definitive Guide, 4th Ed., O'Reilly.

[2] Meier A., M. Kaufmann (2019) SQL & NoSQL Databases: Models, Languages, Consistency Options and Architectures for Big Data, Springer.

[3] Harrison G. (2015) Next Generation Databases: NoSQL, NewSQL, and Big Data, Apress, 256 pp.

[4] Plattner H. (2014) A Course in In-Memory Data Management: The Inner Mechanics of In-Memory Databases, Springer.
[5] Bureva V. (2021) Overview of Big Data Systems, In: Annual of Section "Informatics" of Union of Scientists in Bulgaria, **11**, (in press).
[6] Orozova D., I. Popchev (2020) Cyber-Physical-Social Systems for Big Data, In: Proc. XXI Int. Symp. on Electrical Apparatus and Technologies SIELA 2020, 3–6 June 2020, Burgas, Bulgaria, 334—337.
[7] Orozova D., K. Atanassov (2018) Generalized net model of processes related to Big Data, C. R. Acad. Bulg. Sci., **71**(12), 1679–1686.
[8] Orozova D., K. Atanassov (2019) Model of Big Data MapReduce processing, C. R. Acad. Bulg. Sci., **72**(11), 1537–1545.
[9] Popchev I., D. Orozova (2019) Towards Big Data Analytics in the E-learning Space, Cybern. Inf. Technol., **19**(3), 16–25.
[10] Popchev I., D. Orozova, S. Stoyanov (2019) IoT and Big Data Analytics in E-Learning, In: 2019 Big Data, Knowledge and Control Systems Engineering (BdKCSE), 1–5, doi: 10.1109/BdKCSE48644.2019.9010666.
[11] Angelov P., Y. Manolopoulos, L. Iliadis, A. Roy, M. Vellasco (2017) In: Advances in Big Data: Proc. 2nd INNS Conf. on Big Data, Oct 23–25, 2016, Thessaloniki, Greece, Springer.
[12] Hassanien A., A. Azar, V. Snasael, J. Kacprzyk, J. Abawajy (2015) Big Data in Complex Systems, Springer.
[13] Prabhu C. S., A. Chivukula, A. Mogadala, R. Ghosh, L. Livingston (2019) Big Data Analytics: Systems, Algorithms, Applications, Springer.
[14] Popchev I., D. Orozova (2021) Data Science: Experience and Trends, In: Proc. Int. Conf. on Technics, Technologies and Education ICTTE 2020, J. IOP Conf. Ser.: Materials Science and Engineering, **1031** 012057, IOP Publ., doi:10.1088/1757-899X/1031/1/012057.
[15] Atanassov K. (1987) Generalized index matrices, C. R. Acad. Bulg. Sci., **40**(11), 15–18.
[16] Atanassov K. (2014) Index Matrices: Towards an Augmented Matrix Calculus, Springer, Cham.
[17] Atanassov K. (2018) n-Dimensional extended index matrices, Adv. Stud. Contemp. Math., **28**(2), 245–259.
[18] Atanassov K. (2022) Cartesian products over index matrices, Adv. Stud. Contemp. Math., **32**, (in press).
[19] Atanassov K. (1998) Generalized Nets in Artificial Intelligence, Vol. 1: Generalized Nets and Expert Systems, Sofia, Prof. M. Drinov Academic Publishing House.

*Department of Bioinformatics
and Mathematical Modelling
Institute of Biophysics
and Biomedical Engineering
Bulgarian Academy of Sciences
Akad. G. Bonchev St, Bl. 105
1113 Sofia, Bulgaria
e-mails: k.t.atanassov@gmail.com
krat@bas.bg

**Intelligent Systems Laboratory
"Prof. Asen Zlatarov" University
8010 Burgas, Bulgaria
e-mail: vbureva@btu.bg