Big Data Overview and Connected Research at Óbuda University Alba Regia Technical Faculty

Éva Hajnal
Óbuda University Alba Regia Technical Faculty, Székesfehérvár, Hungary
hajnal.eva@amk.uni-obuda.hu

***Abstract*—21st century is called the data era. In this paper a wide overview is drawn about big data phenomenon. It includes the definition and main characteristics of big data, as well as the conceptual changes that is caused in the thinking and in the way of information gathering, processing and storing. Finally a research example illustrates the big data strategy and as a part of it, the data science methods. The intelligent data analysis of expert database systems is in the focus of this example namely the high education database research. A SOM based prediction model was developed, and its accuracy was investigated on five years real data.**

I. INTRODUCTION

The 21st century is called the data era. Big data means that during the industrial and social processes, in the sensor networks etc. large volume of data is generated with a very large velocity. The processing of these data gives large opportunities and creates interesting challenge to the data scientists as well as to the other sectors.

The computer technics development caused that the cost of the storage of one byte is millionth part less today as it was 15 years ago. The consequence of it the stored data volume started a rapid increasing [1], so by the predictions it will be 41 zettabyte stored data/year till 2020 (Figure 1). By this data explosion started the new data driven paradigm, the name of which is big data. This big data phenomenon has 3, 5 or 7 characteristic Vs [1]–[3](Figure 2). The volume means the amount of data. It can only be handled with help of special technological solutions from the special distributed file storage system (HDFS) to modern NoSQL database types via cloud solutions.

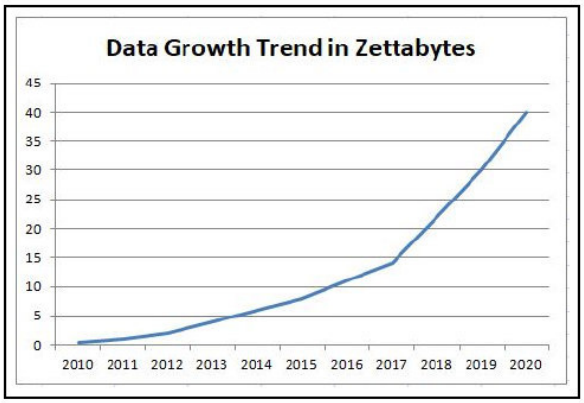

Figure 1 Trend and estimation of stored data volume per year [1].

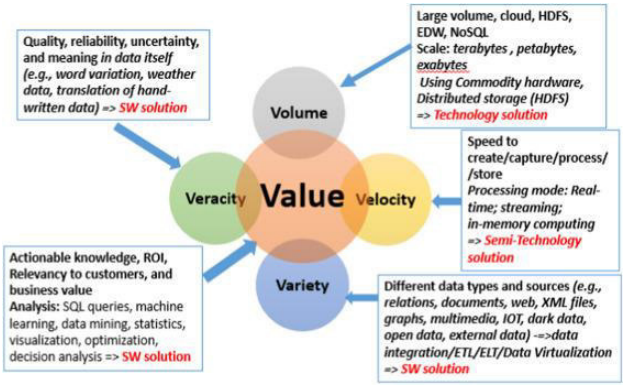

Figure 2 The 5 Vs of Big Data: Volume, Velocity, Variety, Veracity and Value [2]

The velocity of the data creation is large, that means a necessity of the adequate processing and storage. This can be achieved by technological solutions, but additionally needs careful design and appropriate software solutions. The data can be originated from different sources, as well as from internet applications, industrial processes, IoT devices, etc., and later the storage method also can differ and results in the variety of the stored data. In these cases the data integrating software solutions have large importance. Moreover the veracity of data is questionable and has to take care of attention to avoid the misinterpretations. The main feature is the value which can be yielded out from the data by data mining methods. This gives the added value of the big data.

The traditional relational databases after normalization have a performance limit [4], which caused the development of other database formats in this century (Figure 3).

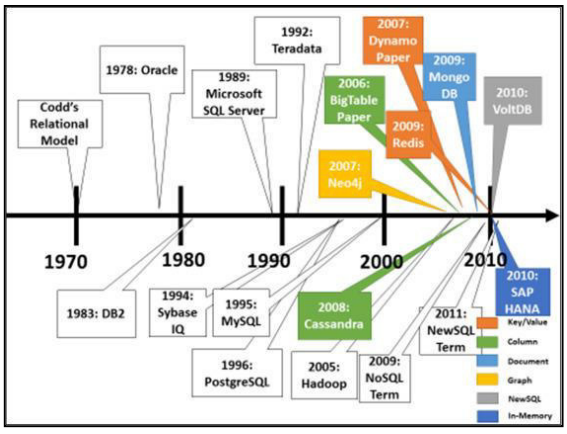

Figure 3 History and evolution of different database systems

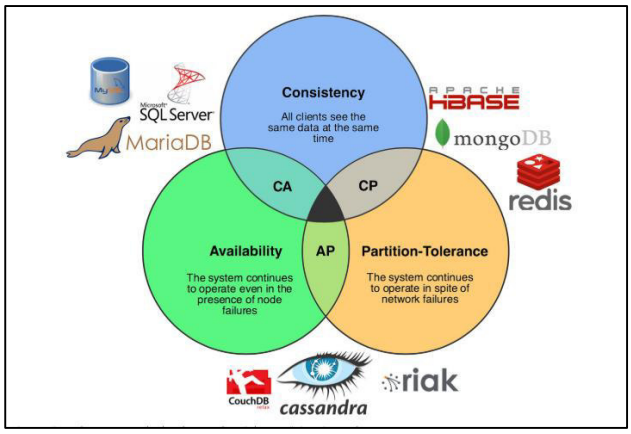

Figure 4 Database types grouped by the Brewer theorem [4].

The large amount of data requires partitioned storage for easy scaling, however traditionally we need to keep our data consistent and available in every moment. To reach this hoped aim is impossible as it was first drew up by Eric Brewer. According to his theorem we have to be satisfied with two from these three requirements and one point of view in choosing the right database system [1]–[4] is the place of it in this system (Figure 4). The lately developed NoSQL databases are in their childhood. According to the demand of them there is a need to develop them into a matured state. The main features of traditional relational databases and these NoSQL systems are compared in table 1. The schema less storage system gives the opportunity of NOSQL databases to process the large amount of continuously raising data. However the missing transaction tracking system makes the traditional relational databases the only acceptable solution in bank systems and other systems where the data tracking is extraordinary important.

The new data driven paradigm changes the point of view, the way of thinking and working of IT specialist, the decision-makers and finally of civils. This new point of view differs from traditional data management in several cases. One important new methodology is the whole dataset sampling, so we collect all data instead of the traditional representative sampling methods. After investigation of these big data we have to accept probability prediction without any reasons. In this new era the large amount of data makes impossible to a human expert to see the relationships between different things. In this situation big data supplemented with AI help can outperform experts [5]. The results of data mining is often in form of correlations, and decision makers have to accept them. This data driven paradigm means that all processes are supported with data analysis including the sentiment analysis as well. For this purpose either of the things can be the initiator of data creation, the name of which is datafication.

II. DATA MINING RESEARCH

During the last ten years a research group was created for data mining investigations. At first the name of the group was Mathematical modelling research team with the leadership of József Lakner and later Éva Hajnal, and this group was reorganized and renamed in 2018.

Table 1 Relational Versus Not Only SQL (NOSQL) Database Management Systems

Relational	NOSQL
Based on relations (normalized sheets)	Variable storage technics (column based, key-value pair, document based, graph databases)
Work with schema	Dynamically varying scheme
Vertically scalable	Horizontally scalable
Querying with standard SQL syntax	Querying with SQL like but not standardized syntax
Complex queries	Basic **C**reate **R**ead **U**pdate **D**elete operations
Transaction tracking	No transaction tracking
Before MySQL 8.0, classic Oracle, PostgreSQL	Apache Cassandra, MongoDB, HBase, Redis, Neo4j

Now it works as intelligent data analytics group. During the period a few research topics were investigated, as well as in basic and applied sciences. Our main partners are Hidrofilt Ltd., University of Pannonia and some multinational firms. The fields and the connected publications are shown in table 2.

Table 2 Scientific research and publications by field

Scientific research	Number of publications
Educational questions	5
Modelling, Cell automaton, ecological modelling	4
Phytoplankton and perifiton research	22
Data mining in economy	4
Handwriting recognition	4
Text classification	2

III. A SOM BASED MODEL

In this paper an intelligent data mining method and its usage is presented. Our purpose was to investigate the financial support and demographic data in Research and Development Sector including the high education. Based on the collected data set our purpose was a prediction with SOM about the expected number of students and the expected scientific results. These calculations were made and published five years ago [6], and now it is the time to investigate the accuracy of that prediction against the real data.

Methods

The Self Organizing Map is an unsupervised learning neural network, which was intended by Kohonen.

These types of neural networks do not get feedback about the output of the data processing. The aim is to find the optimal output based on the input. They consist of only one neuron layer and defines an input and an output space. The output space forms a map. It is important that the algorithm takes into account the places of neurons, consequently in successful data processing can preserve the original topology.

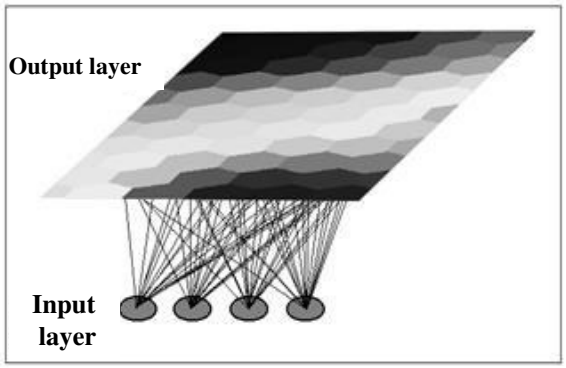

Figure 5 Topology of the Kohonen-type Self Organizing Map

The neighborhoods of neurons are respected during the data processing. The input neurons have weight vectors, which are responsible for the connection between the input and output data.

The steps of the learning algorithm are:

1. Initialization of neural network, number of neurons, weight vectors
2. An input data calculation, distance from every neurons.

$$(d_{ij} = \| x_k - w_{ij} \|), \quad (1)$$

where i and j are row and column indices of a neuron, k is the index of the input data d_{ij} the distance between the input neuron and the i^{th} and j^{th} neuron, x_k is the input data vector and w_{ij} is the weight vector of the i^{th} and j^{th} neuron.

3. The determination of the closest neuron, this is the Best Matching Unit (BMU)

$$w_{winner}(w_{ij} : d_{ij} = \min(d_{mn})), \quad (2)$$

where w_{winner} is the BMU.

4. Actualization of weight vector according to the distance and the learning rate.

$$w_{ij} = w_{ij} + a * h(w_{winner}, w_{ij}) \| x_k - w_{ij} \| \quad (3)$$

where a is the learning rate and h is the neighborhood function.

These 2.-4. steps are iterated till a previously determined criterium is reached.

The Self Organizing Map is an efficient clustering, correlation hunting and estimation method. It is widely used for optical character recognition, for face recognition [7] [8], for robotics to recognize barriers in the environment [9]. In the literature there are examples to scientific [10] and technical modelling [11] usage and financial analysis. However only few examples deal with data estimation or prediction with the help of SOM.

The theoretical basis of the estimation is to find the best matching units of an input data point, the calculation of the mean weight vector and denormalization.

During calculations standard setup (Euclidean distance, Gauss neighborhood functions, standard batch training algorithm, hexagonal lattice) and range type data normalization were used.

Data

Digitalized Data: Data were downloaded from the official website of the Hungarian Central Statistical Office: main ratios of R+D (1990-2012), staff number in R+D (1990-2012), financial sources of the R+D input (2000-2012), number of publications (1990-2012), R+D support (1990-2012), patent activities (2000-2012).

Paper based data: The R&D data before 1990 came from the Statistical Yearbooks.

Results and discussion

The collected dataset was validated and then some statistical calculations were made. Among them there were traditional statistical steps and the previously described SOM based model, which was used not only for correlation hunting, but moreover for prediction. Based on this model a 20 years long prediction was made to the total number of students in high education [6]. This prediction is presented on figure 5. The prediction seemed very pessimistic and thus got some critics. Additionally it is not based on widespread surveys, only on statistics. Consequently it had only statistical base and no detailed reasoning by each factors and effects.

After five years the first check of this prediction was executed against the Educational Office's data [12]. The absolute difference between the model and real data was calculated and presented in Table 3. The average accuracy of our model was -0,36%. The calculated prognostic accuracies that based on the predicted yearly changes and their difference from real yearly changes are presented in Table 4. In this system the average accuracy of prediction is 17% which is yet acceptable. However its changes are between -23% and 51% which in connection with the quite exact absolute accuracy means a bit of difference (1 year) in the temporal pattern.

Table 3 The accuracy of our five year old prediction [12][6], [13]–[15]

Years	Statistical number	Estimated number	Difference	Accuracy
2013	320 124	325 624	5500	1,7%
2014	306 524	304 362	-2161	-0,7%
2015	295 316	292 874	-2441	-0,8%
2016	287 018	286 122	-895	-0,3%
2017	283 350	278 617	-4733	-1,7%

It is concluded, that by the first five years the model is quite accurate and seems to be relevant in prediction. It is based on only few factors, so we interestingly wait its later accuracy. Generally the models are differing increasingly by the time. The social trends and social processes based on the behavior of a large multitude and their inertia cause a good predictability in social processes which was demonstrated by this example.

Table 4 Accuracy of the predicted yearly changes

Year	Real change	Estimated change	Difference	Accuracy
2013-14	-21262	-13600	-7662	36%
2014-15	-11488	-11208	-280	2%
2015-16	-6752	-8298	1546	-23%
2016-17	-7506	-3668	-3838	51%
				17%

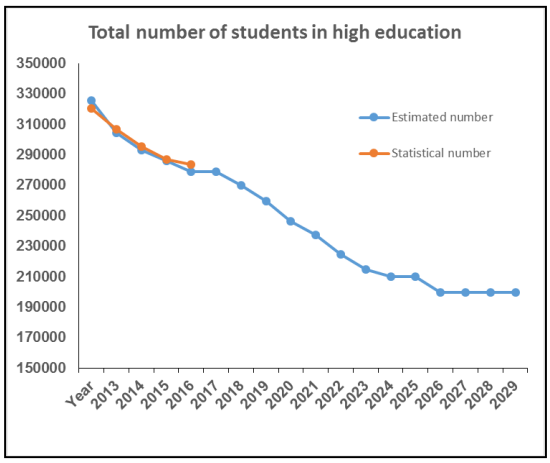

Figure 5 Accuracy of the prediction by five year statistics

IV. Conclusion

In this paper the big data principle was defined and its features were presented. The so called data driven paradigm changes our life, the way of working and thinking. It brings new technological solutions as well as in file storage and handling system, in database management and cloud computing. These solutions often are in their childhood, and we are waiting for new and better solutions to take the IT work easier and the systems safer. It is clear that conceptual thinking is generally behind the technology. It means the missing concepts in database design especially in design of hybrid SQL-NOSQL systems. More interesting are the changes of our thinking, the consequence of the data mining and statistical point of view. An interesting example demonstrated the acceptance of statistical analysis and correlation based prediction.

Acknowledgment

I would like to thank the help and the effort of colleagues and students. József Lakner, Valéria Varga, Péter Ivanics, Balázs Nagy, Csaba Ambruzs, Gaye Ediboglu Bartos and the colleagues of the Limnological Department of University of Pannónia helped our work.

References

[1] S. Muhammad and F. Akhtar, *Big Data Architect ' s Handbook*. Packt Publishing Ltd., 2018.

[2] V. C. Storey and I. Song, "Data & Knowledge Engineering Big data technologies and Management : What conceptual modeling can do," *Data Knowl. Eng.*, vol. 108, no. February, pp. 50–67, 2017.

[3] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," *J. Bus. Res.*, vol. 70, pp. 263–286, 2017.

[4] J. R. Lourenço, "Choosing the right NoSQL database for the job : a quality attribute evaluation," *J. Big Data*, pp. 1–26, 2015.

[5] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, and L. Sifre, "Article Mastering the game of Go without human knowledge," *Nat. Publ. Gr.*, vol. 550, no. 7676, pp. 354–359, 2017.

[6] M. Majdán and B. Nagy, "Self Organizing Maps in Economical Analysis (Önszervező térkép (SOM) használata gazdasági elemzésekben)[in Hungarian]," Óbuda University, 2013.

[7] C. H. Lee, D. S. Seong, and K. H. Park, "Face recognition using self-organizing map," *J. Korea Inf. Sci. Soc.*, vol. 20, no. 11, pp. 1730–1738, Nov. 1993.

[8] J. L. Alba, a. Pujol, and J. J. Villanueva, "Separating geometry from texture to improve face analysis," in *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*, 2001, vol. 2, pp. 673–676.

[9] K. Ishii, S. Nishida, and T. Ura, "A self-organizing map based navigation system for an underwater robot," in *2004 IEEE International Conference on Robotics and Automation IEEE Vol. 5*, 2004, p. 5726.

[10] A. Ultsch and R. Frank, "Self-organizing feature maps predicting sea levels," vol. 144, pp. 91–125, 2002.

[11] M. Cottrell, P. Gaubert, C. Eloy, D. François, G. Hallaux, J. Lacaille, and M. Verleysen, "Fault prediction in aircraft engines using Self-Organizing Maps," 2009.

[12] OH, "Statistics." [Online]. Available: https://www.oktatas.hu/felsooktatas/kozerdeku_adatok/felsooktatasi_adatok_kozzetetele/felsooktatasi_statisztikak. [Accessed: 15-Sep-2018].

[13] B. Nagy, M. Majdán, V. Varga, É. Hajnal, N. Balázs, M. Márk, V. Valéria, and N. H. Éva, "The Use of Self-organizing Maps (SOM) in Demographic Analysis," in *10th International Symposium on Applied Informatics and Related Areas (AIS 2015): AIS 2015*, Székesfehérvár: Óbudai Egyetem, 2015, pp. 22–26.

[14] E. Hajnal, B. Nagy, and V. Varga, "Intelligent estimation method for demographic effect onto the research and development sector in Hungary," in *SACI 2016 - 11th IEEE International Symposium on Applied Computational Intelligence and Informatics, Proceedings*, 2016.

[15] B. Nagy, M. Majdán, V. Varga, and É. Hajnal, "The Use of Self-organizing Maps (SOM) in Demographic Analysis," in *10th International Symposium on Applied Informatics and Related Areas (AIS 2015): AIS 2015*, O. Tamás, Ed. 2015.