Perspective

iMedPub Journals
www.imedpub.com

DOI: 10.21767/2572-5610.100051

Insights in Biomedicine
ISSN 2572-5610

2018
Vol.3 No.3:16

Extracting Actional Information from the Heterogeneity of Big Data

Berkovich S*

Department of Computer Science, The George Washington University, Washington, D.C., USA

Received: December 25, 2018; Accepted: December 28, 2018; Published: December 31, 2018

*Corresponding author:
Dr. Simon Berkovich

berkov@gwu.edu

Professor Emeritus of Engineering and Applied Science, Department of Computer Science, The George Washington University, Washington, D.C. 20052, USA.

Tel: 202 994 8248

Citation: Berkovich S (2018) Extracting Actional Information from the Heterogeneity of Big Data. Insights Biomed Vol.3 No.3:16

Perspective

Traditional way of thought to get knowledge for making decisions under problematical circumstances, especially, in convoluted biomedical situations, is to collect as much information as possible. Such an approach takes for granted that the more information could be collected the more successful this tactic should be. So, it comes as a surprise that notwithstanding the colossal efforts and remarkable technological advances this obvious "Big Data" approach does not bring in the expected results [1].

The problem lies not in the very large size of the considered systems. Analysis of any amounts of loosely structured and diversified information is a very nontrivial task per se; and big sizes urge a qualitatively different approach. This analysis needs two separate steps: First, selecting a relevant subset of data approximately matching the object of study, and second, determining which attributes of the objects are actually responsible for their specific properties. As well known, the initial first step of choosing data set is to provide an approximate description of the investigated object. This involves heavy procedures for searching and clustering [2]. Interestingly, nowadays, a well-known parable "Knowledge is Power" has an astounding twist: As long as complex searches commonly utilize brute force computers consume excessive amount of power, about 5% of total energy consumption [3].

We have found a completely novel solution for the vital initial data selection obviating the ubiquitous insurmountable requirements of searching and clustering. This invented selection of a relevant set of information items from Big Data repository has been presented in numerous publications, as it has been investigated in different aspects for several years by a group of doctoral students of the GWU. A general consideration of these publications is given in a book chapter and our short review is presented [4,5].

The suggested procedure is based on an exceptional feature of the so-called perfect error-correction Golay codes that can be applied to partition the binary cube of the 23-bit vectors displaying the attributes of given information items [6]. Presenting the attributes by means of a 23-bit template furnishes an assortment of 12-bit indices, which provide fault-tolerance facilities for fuzzy matching and retrieval [7].

For the consideration of amorphous data, we use what we call "Meta Knowledge 23-bit Templates" [8]. Namely, in specification of certain categories of knowledge we introduce sets of 23 inquiries by 23-bit patterns. There is a well-known amusement game called "20 questions". In this game a person thinks of a certain concept, and other people try to guess what this concept is by posing no more than 20 questions. So, a 23-bit template should be sufficient to produce a reasonable characterization for different information items. The organization of the suggested process requires establishing a set of "Metaknowledge 23-bit Templates". These metaknowledges will present an additional intrinsic component of computer language tools, such as dictionaries, thesauruses, etc.

We introduce a novel extraordinary type of operation to select appropriate information items for the Big Data analysis-memory cluster access. Ordinarily, such a selection begins with choosing information items from a predetermined size neighborhood of a given request as Hamming Distance in associative memory. Yet, in this neighborhood certain information items although close to the request will be still far away from each other. So, the functional relationships could not be simply revealed without additional separation of the selected data. The memory cluster access intrinsically provides an automatic grouping of relevant data. This novel operation becomes feasible due to the suggested application of the pigeon-hole principle to the Golay Code technique [9]. This work was awarded the first prize at the GWU Research Showcase 2014 [10].

A certain possibility of using the presented clustering methodology was considered [11]. For a conclusive second step to actually obtain a particular required knowledge we need an efficient algorithm for extracting the functional combinations of the attributes from the indicated ensembles of clusters, e.g., biomarkers set for a particular disease These decisive analytics work essentially benefits from our memory cluster access. The Big Data problems are not much determined by the bigness of the information collections, but also by their amorphness urging for qualitatively new design of information processing algorithms. Thus, we introduce a novel system ABC (Amorphous Block Clustering), which seems to find out a unique approach to the Big Data problems. Precision Medicine cannot successfully advance without effective tools for Big Data explorations.

References

1 https://www.newsweek.com/2014/08/01/scientists-question-big-price-tags-big-data-260690.html.

2 Dunham MH (2003) Data mining: Introductory and advanced topics. Prentice Hall, Pearson Education Inc, Upper Saddle River, New Jersey, USA.

3 Page ML (2018) Knowledge means power. New Scientist 240: 22-23.

4 Duo L (2016) A qualitatively different principle for the organization of big data processing, Big Data: Storage, sharing, and security, CRC Press, UK.

5 Berkovich S (2017) On the big data application to the practice and theory of biomedicine. Adv Biochem Biotehcnol: ABIO p: 145.

6 Berkovich SY, El-Qawasmeh E (2000) Reversing the error-correction scheme for a fault-tolerant indexing. Comp J 43: 54-64.

7 Berkovich E (2007) Method of and system for searching a data dictionary with fault tolerant indexing. US, Patent No.: 7,168,025.

8 Bari N, Vichr R, Kowsari K, Berkovich SY (2014) 23-bit metaknowledge template towards big data knowledge discovery and management. Int Conf Data Sci Adv Analyt (DSAA), USA.

9 Yammahi M, Kowsari K, Shen C, Berkovich S (2014) An efficient technique for searching very large files with fuzzy criteria using the pigeonhole principle. Fifth Int Conf Comput Geospat Res Appl 2: 1.

10 https://gwtoday.gwu.edu/researchers-showcase-technologies-commercial-potential.

11 Alsaby F, Alnowaiser K, Berkovich S (2015) Golay code transformations for ensemble clustering in application to medical diagnostics. Int J Adv Comput Sci Appl 6: 1-2.