International Journal of Control and Automation
Vol. 9, No. 4 (2016), pp. 233-244
http://dx.doi.org/10.14257/ijca.2016.9.4.23

A Big Data Analytics based on Multi-dimensional Matrix for Large Text Datasets

Fan Linxiu

School of Mathematics and Computer Science; Gannan Normal University, Ganzhou, China
feelingshow@126.com

Abstract

Big Data is becoming more and more significant these years since our daily life is facing huge number of data as the millions of electronic devices. Big Data is not only with the huge volume or size, but also with the high complexity. This paper presents a multi-dimensional matrix model for analyzing the large text datasets based on the attributes, which come from the key words from the texts. These key words form an N dimensional space. Thus, the individual information could be presented by an M×N matrix. The multi-dimensional matrix approach has been compared with GA and PSO algorithm so as to test the efficiency and effectiveness of different approaches on analyzing the text datasets. From the experiments, it is observed that the proposed approach outperforms GA and PSO in sufficiency and computational cost. Some key findings are: For high dimensional Big Text Data, at the beginning, PSO has the best sufficiency from 0 to 10. After that, from 10 to 1000, the prosed multi-dimensional matrix approach significantly outperforms GA and PSO. For Connect-4 data samples, the time cost of proposed approach is only 352153.6 unit of time, while GA takes 613601.4 which is more of about half the time cost and PSO takes 469464.1.

Keywords*: Graphic Theory, Data Mining, RFID; Sensor Network, Apriori Algorithm*

1. Introduction

Big Data is becoming very hot these years since our daily life is facing huge number of data as the millions of electronic devices, Internet usage, social behaviors, and natural discovery activities are carried out [1]. Traditional database or processing techniques cannot manage or deal with such enormous datasets efficiently and effectively [2-4]. For example, based on the Internet application, large quantity of texts is generated from Facebook, Twitter, and communication app like Whatsapp, Line, Wechat, *etc.* Such data is sharply increasing every second. It was reported that in the year of 2006, the data generated from individuals reached TB and the whole data size is about 180EB globally [5]. In the year 2011, the data reached 1.8ZB and it is estimated that the data volume will be 35.2ZB which is 44 times than those in 2011 [6].

Big Data is not only with the huge volume or size, but also with the high complexity. Big Data always includes the transaction and interactive data which have large scale of complexity [7]. Thus, the traditional methods are not able to capture, manage, and process these big datasets. The Big Data may be created from main three perspectives: (1) Enormous transaction data. From the ERP (Enterprise Resource Planning) and data warehouse application in the online transaction processing and analyzing system, traditional relation data, non-formatted and semi-constructed data are still increasing quickly. As more data and transaction shifted to public or private cloud application, such data will become more and more sophisticated. For example, the internal transaction information including online transaction and analysis data is formatted. Through the database, users can access these data through managing the static historical data. By

ISSN: 2005-4297 IJCA

means of that, what happened in the past could be explored using some analytics approaches. (2) Massive interactive data. These data come from great myriad of social media such as Facebook, Twitter, LinkedIn and mobile communication Apps [2]. Such data include the call detail records, equipment and sensor information, GPS and location mapping data, manage file transfer protocol and its huge figure/image files, web texts and click stream data, e-mails, and so on. By full use of such data, we can predict what will be happened in the near future. (3) Big Data processing. Light databases are used for receiving the data from various clients. After that, all the data will be transferred to a central huge distributed database or storage sets [6]. The distributed database is able to query and classify the enormous data so that general analysis on the data could be achieved. Meanwhile, using the data mining technology, advanced usage of the data could also be realized. For example, YunTale is a new distributed database which are based on the traditional database technique twining with NoSQL [8-9]. It is able to build up a PB database to manage the huge datasets through cooperating hundreds of distributed storage devices.

As the US government announced a plan on 29 March 2012 that 0.2 billion US dollars will be invested for promoting the Big Data related industry which made a "Big Data Strategy" in the whole nation [10]. Thus, Big Data has been widely studied, investigated, and implemented worldwide. Text-based Big Data is one of the hottest topics since the Internet applications are used daily by large quantity of people globally. Elder et al discussed the text mining using the statistical analysis for non-structured text data applications from the view of practical aspects [11]. For dealing with the family psychology, Atkins *et al* analyzed the text data applying a coding system, which quantifies the text data [12, 13] described a process of building statistical models of the Slovak language with large vocabulary trained on the text data gathered mainly from Internet sources with several smoothing techniques for different sizes of vocabulary have been used in order to obtain an optimal model of the Slovak language.

Besides all the efforts, there are still some research gaps need to fulfill. Firstly, the traditional statistics methods are still time-consuming. Simple analysis of frequency and deviations are inadequate for advanced decision-makings [14]. Secondly, the texts datasets are with very diversity. Data mining approaches such as decision tree and neural network are basic for extracting the information and knowledge from large number of texts [15-18]. Additionally, when face enormous text datasets, it is over the computational time when getting the results. This paper presents a multi-dimensional matrix model for analyzing the large text datasets. This Big text Data analytics is based on the attributes, which come from the key words from the texts. These key words form an N dimensional space. Thus, the individual information could be presented by an M×N matrix.

The rest of this paper is as follows. Section 2 presents the problem description which is based on the online questionnaire system. Section 3 illustrates the characteristics of the defined multi-dimensional matrix and its application with hypergraph. Section 4 reports on the experiments on graphical texts by applying the multi-dimensional matrix. Conclusions and future work are presented in Section 5.

2. Problem Description

The description of problem is based on the attributes which are related to the questions and key words from the online questionnaire. The attribute values are related to the options and key attitude. Each attribute is defined as a dimension in a space. Thus, a N dimensional space could be formed. Each surveyor uses a M×N matrix (M is the maximum number of options). Each option in a dimension is an element, which could be a vector or a matrix. When it is selected, '1' is presented. Otherwise, '0' is presented. Each surveyor's information may include N elements so that all the

questionnaires with the feedback from the surveyors could be collected.

Table 1. Demonstrative Examples

Options	Type 1	Type 2	Type 3								
A	0	1	0	0	1	0	0	0	0	0	0
B	1	0	0	1	0	0	0	0	0	0	0
C	0	1	1	0	0	0	0	0	0	0	0
D	0	0	0	0	0	1	0	0	0	0	0
E	0	1	0	0	0	0	1	0	0	0	0
F	0	0	0	0	0	0	0	1	0	0	0
G	0	0	0	0	0	0	0	0	0	0	1
H	0	0	0	0	0	0	0	0	0	1	0
I	0	0	0	0	0	0	0	0	1	0	0

A demonstrative example is presented in Table 1 for illustrating the problems. There are three types of questionnaires and the selected options are marked with '1'. Type 1 is single choice. Type 2 is multiply choices and Type 3 is sequencing questions. The option for type 1 is 'B'; type 2 is 'ACE' which should be presented by a vector and type 3 is 'CBADEFIHG' which needs increasing of the dimension with a matrix. Assume that there are N questions whose information could be used by a $M \times N_1$ matrix ($N_1 \geq N$). K surveyor's N questions' information could be presented by K $M \times N_1$ matrix. The data from different data sources could be converted and presented. After forming several multi-dimensional matrix, the data could be integrated. The data integration combines all the rectangles so as to form an nD matrix according to some rules. The nD matrix is presented as Figure 1.

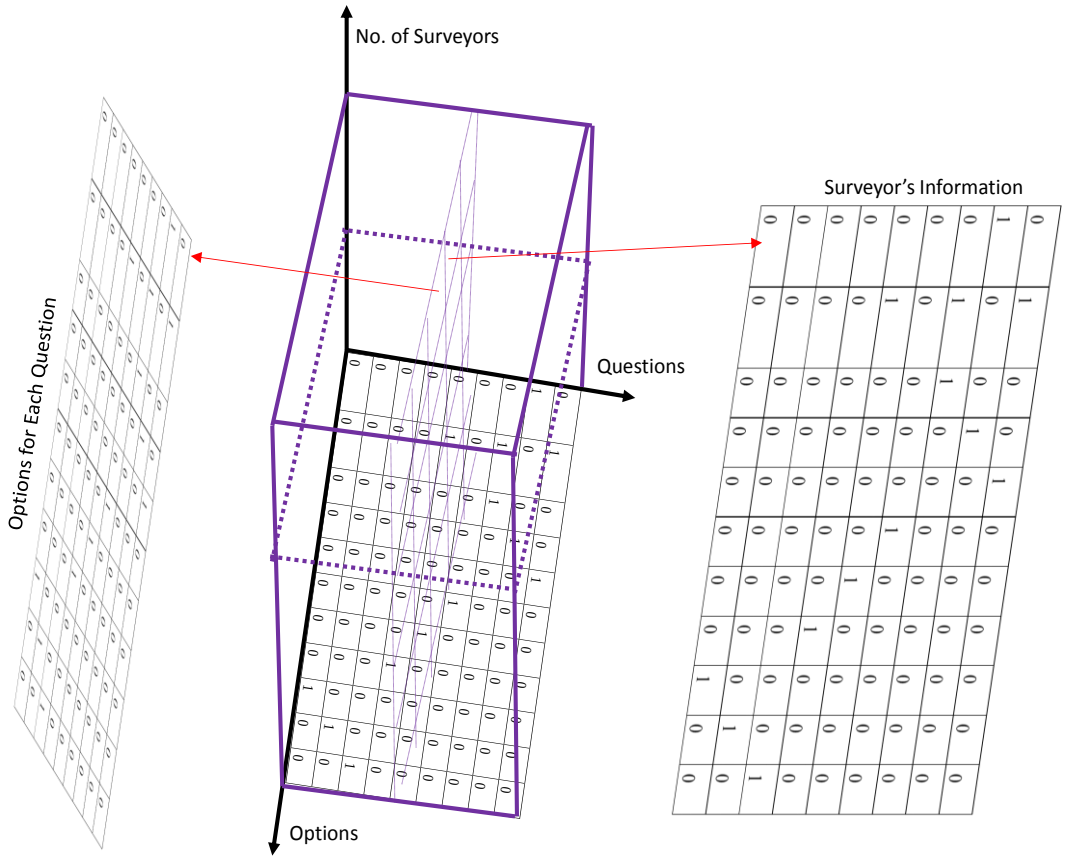

Figure 1. Principle of Multi-dimensional Matrix Model

From Figure 1, the multi-dimensional matrix model has three axis, which are questions, options, and number of surveyors. Questions axis takes a question and a key word as a unit, which is associated with the width of rectangle. The axis of surveyor amount takes each individual as a unit, which is associated with the height of rectangle. The option axis will be A~F, which is the quantity of options from questionnaires. It is associated with the thickness of the rectangle. For the information extraction, each section plane of height could get the information about each questionnaire or surveyor. While, each section plane of width may obtain the each option or each attribute information. From different sections, the multi-dimensional matrix can be extended by calculating or operating the section plane.

3. Multi-dimensional Matrix and its Applications

3.1. Several Definitions

There are several definitions when summarizing its characteristics.

Definition 1: 3D matrix. Define a 3D matrix is consisted by $n \times p \times q$ three dimensional array, where n presents the layers' amount. $p \times q$ implies that each layer is a $p \times q$ matrix. $X_{i,j,k}$ is a i row and j column in layer k of a 3D matrix X . n, p, q indicates the height, thickness, and width of the 3D matrix. Relating to the demonstrative example mentioned previously, the height n presents the

surveyor's quantity, thickness p means the maximum number of options in a question. Width q indicates the attributes (questions).

Definition 2: M-matrix. M-matrix is a specific matrix where some elements are '1' and others are '0'. If all the elements are '1', the $n \times p \times q$ matrix is $M^{(1)}$-matrix. And $M^{(0)}$-matrix means all the elements are '0'.

Definition 3: Left Multiplication. Assume that A is a $r \times p$ matrix, X is a $n \times p \times q$ matrix. The left multiplication of A and X is carried out through A left multiple the matrix in each layer in X. Then, the n resulted matrix is sequenced by the order from X. The dimension is defined as $\underset{n \times r \times q}{Y} = \underset{r \times p}{A}\ \underset{n \times p \times q}{X}$.

$$Y_{i,j,k} = \sum_{s=1}^{p} a_{i,s} x_{s,j,k} \tag{1}$$

When A is a $1 \times p$ vector, X is a $n \times p \times q$ matrix. Quadratic AXA' is a n dimensional vector:

$$AXA' = (AX_1A' \quad AX_2A' \quad \ldots \quad AX_nA')$$

$$= (\sum_{i=1}^{p}\sum_{j=1}^{p} x_{i,j,1} a_i a_j \quad \sum_{i=1}^{p}\sum_{j=1}^{p} x_{i,j,2} a_i a_j \quad \ldots \quad \sum_{i=1}^{p}\sum_{j=1}^{p} x_{i,j,n} a_i a_j) \tag{2}$$

Definition 4: [] multiplication. Assume that A is a $m \times n$ matrix, X is a $n \times p \times q$ matrix. $Y = [A][X]$ means A multiple each layer in X:

$y_{i,j,s} = \sum_{k=1}^{n} a_{s,k} x_{i,j,k}$ with the dimension $\underset{m \times p \times q}{Y} = \underset{m \times n}{[A]}\ \underset{n \times p \times q}{[X]}$.

According to the definition 4, we can get the following operations: (g, g_1, g_2 are numbers like integer or float)

(1) $[\lambda A][X] = [A][\lambda X] = \lambda [A][X]$

(2) $[A + B][X] = [A][X] + [B][X]$

(3) $[A][X] = [X][A']$

(4) $[AB][X] = [A][[B][X]]$

(5) $(g_1 + g_2)X = g_1 X + g_2 X$

(6) $g(X + A) = gX + gA$

(7) $g_1(g_2 X) = (g_1 g_2)X$

Definition 5. Multiplication of multi-dimensional matrix. Let $A = [a_{i,j,k}]_{p,q,r}$, $B = [a_{j,k,l}]_{q,r,s}$, $C = [c_{i,j,l}]_{p,q,s}$, $D = [d_{i,k,l}]_{p,r,s}$ where

$$c_{i,j,l} = \sum_{k=1}^{r} a_{i,j,k} b_{j,k,l} \quad i = 1, 2, \ldots, p;\ j = 1, 2, \ldots, q;\ l = 1, 2, \ldots, s \tag{3}$$

$$d_{i,k,l} = \sum_{j=1}^{q} a_{i,j,k} b_{j,k,l} \quad i = 1, 2, \ldots, p;\ k = 1, 2, \ldots, r;\ l = 1, 2, \ldots, s \tag{4}$$

(3) is the R-M operation of A and B, which expressed as $C = A_{*RM} B$.

(4) is the L-M operation of A and B, which expressed as $D = A_{*LM} B$.

3.2. Characteristics of Multi-dimensional Matrix

According to the above definitions, the characteristics of multi-dimensional matrix could be based on the real-life significance and requirements of the text analysis from Big Data. Such characteristics could be expressed as follows.

Firstly, each matrix of cross section of the multi-dimensional presents all the information of the questions in the questionnaires.

Secondly, each vertical section is a matrix, presenting all the information of each question.

Thirdly, based on the left multiplication (definition 3), the weight of each question could be determined. Thus, the multi-dimensional matrix could be converted to traditional matrix operations.

For example, when the left multiplication of $A_{1\times p}$ and $X_{n\times p\times q}$ is carried out, where sector A is the weight of questions, X is the multi-dimensional matrix. After the multiplication, a $n \times q$ matrix will be generated, in where each row presents the information of a surveyor and each column presents the importance after calculating the weights.

$$Y=AX=(5,\ 1,\ 0)\begin{vmatrix} 1-0-0-1 \\ 0-1-1-0 \\ 0-0-0-0 \end{vmatrix}\begin{vmatrix} 0-0-0-0 \\ 0-0-1-1 \\ 1-1-0-0 \end{vmatrix}\begin{vmatrix} 0-0-1-0 \\ 1-1-0-1 \\ 0-0-0-0 \end{vmatrix} = \begin{bmatrix} 5 & 1 & 1 & 5 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 5 & 1 \end{bmatrix}$$

Assume that $A = (5,1,0)$, presenting the importance (weight) of three options in a question. X includes three surveyors whose options in four questions. The importance matrix is $Y = AX$.

3.3. A Typical Application on Image Big Data Processing

The resulted matrix could be applied in image processing with large number of data. For simplicity, we take the above example as a demonstrative case for the Big Data analytics on image. The multi-dimensional matrix could relate to the pixel of the images with large text datasets. In the matrix, the elements '1' could be replaced by a black pixel, while the '0' is blank (as white). After getting the pixel matrix, the following operations could be carried out for Big Data analytics:

1. The projection operation will be implemented by the section n. A plane with large number of pixels could be obtained. The pixel values are Big Data image information which has high density since great myriad of matrix will be sequenced layers by layers.
2. The pixel value is determined by the density of the text in the matrix. That means the frequency of the text. Different gray values are used for presenting the pixels. The gray values indicates the density of each point. The points could be regarded as the items of associated rules.
3. The hypergraph analysis could be used on the big image data. A hypergraph

is a generalization of a graph in which an edge can connect any number of vertices. Then the matrix in each layer can connect with another one through vertical connection. Such connection may reflect the association or relationship from large number of pixels. For example, the GIS remote sensing image with oceans or mountains.

4. Using a hyperedge which connects several attributes (pixels) for presenting the data, the big image data could be projected from top to down. The density of the projection implies the density of hyperedge. Thus, the hypergraph principle will reveal the hidden information from large number of text data from a big data image.

4. Experiments and Discussions

The multi-dimensional matrix is tested by the large text datasets for a Big Data analytics so as to validate the feasibility and practicality of the proposed methodology. Four groups of text datasets are used for this purpose. Table 2 shows the four group text data with different dimensions.

Table 2. Experiment Datasets

Database	Datasets	Data Volume	Data Samples	Dimensions
UCI	Chess	1.17T	28056	6
	Solar Flare	21.6T	13896	10
	Connect-4	22.07T	67557	42
	Plants	61.2T	22632	72

The experiment environment is as follows: Computer with CPU Intel(R) Core™ i7-3770 @ 3.40GHz, RAM 16.0GB, system type: 64-bit Operating System with Windows 7 Professional.

Due to the characteristics of multi-dimensional matrix, we take the connect-4 datasets for example how the experiments are carried out and the text big data from online questionnaires could be analyzed. Firstly, due to the large dimensions of text data, projection operation has been used for reducing the dimension and establishing the hypergraph. Second, after reducing the dimension, the weight of each questions or options could be adjusted. Thirdly, the projection operation on surveyor dimension could be implemented. Thus, the amount of surveyors on each options could be obtained. Fourthly, a hypergraph with different points is formed. The hypergraph's points have different density. Figure 2 shows an example of the text connection with different dimensions at different support and confidence coefficient after the several steps processing.

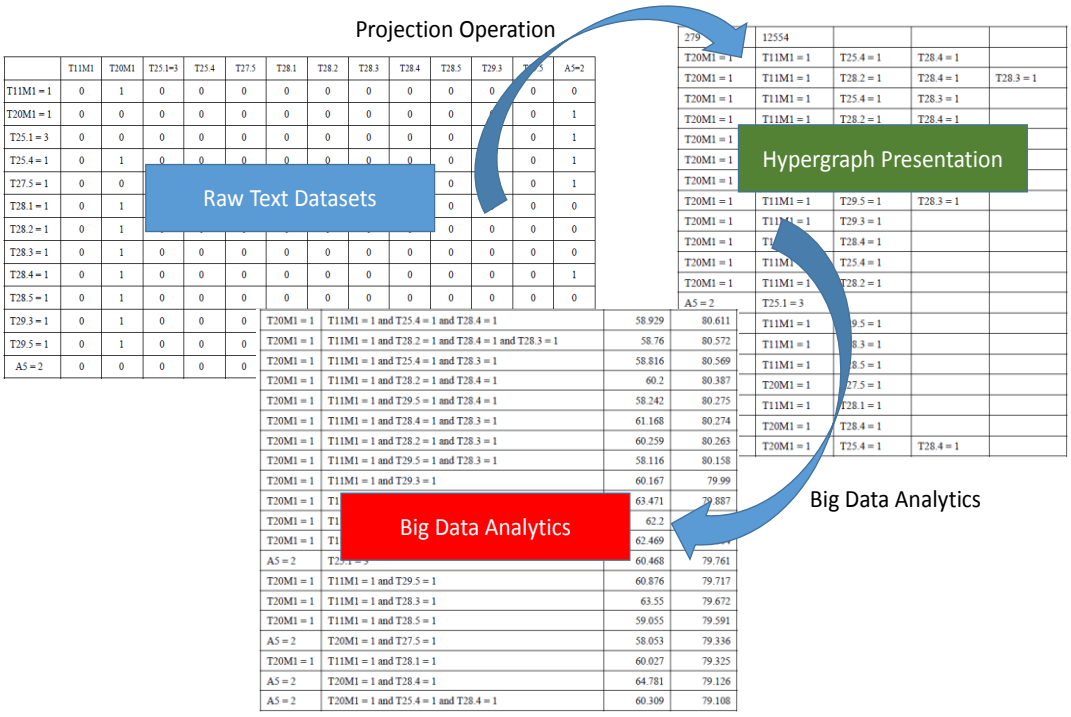

Figure 2. A Demonstrative Example of Processing

The multi-dimensional matrix approach has been compared with GA (Genetic Algorithm) and PSO (Partial Swarm Optimization) algorithm so as to test the efficiency and effectiveness of different approaches on analyzing the text datasets. Figure 3 presents the experimental results from the above four groups of datasets.

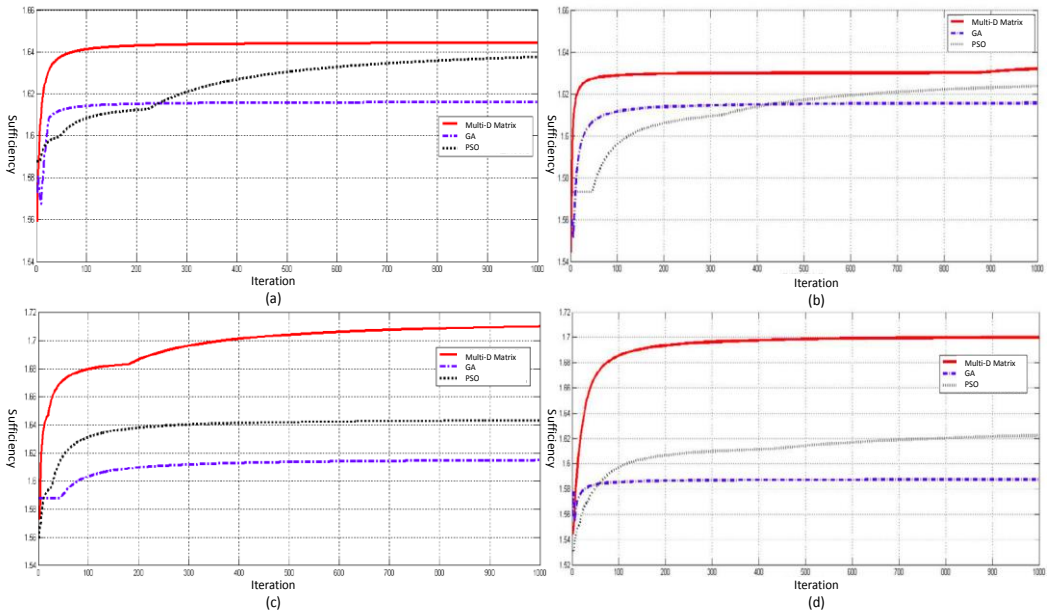

Figure 3. Experimental Results

The three methods implemented on Chess and Solar Flare datasets will present different movement of the sufficiency as the iteration increases. Figure 3 (a) and (b) show the trends of sufficiency. As the dimensions of both datasets are 6 and 10, which are low dimensional dataset. As shown in the beginning, the sufficiency of multi-dimensional

matrix approach is lower than those of GA and PSO. At this stage, PSO has the best performance due to its optimization by having a population of candidate solutions, here dubbed particles, and moving these particles around in the search-space according to simple mathematical formulae over the particle's position and velocity. However, as the iteration increases, the proposed approach outperforms the other two. After 20 times iteration, the multi-dimensional matrix has the better performance in getting the sufficiency on both datasets.

Figure 3 (a) shows that the PSO has a bigger change than the other two approaches. At the first stage, the sufficiency increases sharply as the iteration from 0 to 220. After 220, the ratio increases stably. While, the proposed approach in this paper and GA are sharply increased at the very beginning. After that, the sufficiency is without change as the iteration increases.

Figure 3 (b) shows the same trends of multi-dimensional matrix approach and GA. But the PSO has a little difference. At the beginning, when the iteration from 0 to 50, the sufficiency from PSO is with no change. However, from 50 to 400, the sufficiency is sharply increased. Both (a) and (b) indicate that the GA is the worst approach that because GA takes more steps to carry out the analysis such as coding, decoding, crossover, *etc.*

Figure 3 (c) and (d) show the results from Connect-4 and Plants which have the dimensions 42 and 72, which are high dimensions. From the diagram, the trends from both (c) and (d) are similar. At the beginning, PSO has the best sufficiency from 0 to 10. After that, from 10 to 1000, the prosed multi-dimensional matrix approach significantly outperforms GA and PSO. That implies the proposed approach is capable of dealing with high dimension text datasets and get better analytics results.

Table 3. Computational Results

Approach	Chess	Solar Flare	Connect-4	Plants
Multi-D Matrix	1322.2	1196.7	352153.6	671769.7
GA	1852.5	1977.8	613601.4	898526.5
PSO	1658.3	1587.6	469464.1	763765.8

The computational costs are also examined in the experiments. Table 3 shows the time costs on the above data samples. It could be observed that the proposed approach has better computational cost than GA and PSO. For low dimension text dataset, GA has the worst performance on computational costs. That means, GA has to take more time to get the results. As the increasing of dimensions, the proposed approach has obvious advantages on computational costs. Take Connect-4 for example, the dimension is 42 which is a high dimensional big text datasets. The time cost is only 352153.6 unit of time, while GA takes 613601.4 which is more of about half the time cost and PSO takes 469464.1.

5. Summary

This paper introduces a Big Data analytics approach on large text datasets using a multi-dimensional matrix principle. The problem description is based on an online questionnaire system which carries out the survey. Several definitions are proposed for highlighting the characteristics of multi-dimensional matrix. Experiments have been implemented to validate the feasibility and practicality by comparing the proposed approach with GA and PSO.

Future research directions will be carried out from several aspects. Firstly, the matrix operation with different weights should be improved. In any survey system, AHP

(Analytic Hierarchy Process) could be used since AHP can analyze complex decisions based on mathematics and psychology. Secondly, the Big Data analytics approach could be extended into other application fields such as image processing, where image could be converted into texts. Additionally, Internet text from Facebook, Twitter, and Instagram could be carried by the proposed approach for mining more information and knowledge for analyzing different individual behaviors and habits. Thus, potential market margin could be explored.

References

[1] D. Zeng and R. Lusch, "Big Data Analytics: Perspective Shifting from Transactions to Ecosystems". IEEE Intelligent Systems, vol. 28, no. 2, **(2013),** pp. 2-5.
[2] R. Y. Zhong, G. Q. Huang, S. L. Lan, Q. Y. Dai, C. Xu, and T. Zhang, "A Big Data Approach for Logistics Trajectory Discovery from RFID-enabled Production Data," International Journal of Production Economics, vol. 165, **(2015)**, pp. 260-272.
[3] R. Y. Zhong, Q. Y. Dai, T. Qu, G. J. Hu, and G. Q. Huang, "RFID-enabled Real-time Manufacturing Execution System for Mass-customization Production". Robotics and Computer-Integrated Manufacturing, vol. 29, no. 2, **(2013)**, pp. 283-292.
[4] S. Wilkes, "Some impacts of big data on usability practice". Communication Design Quarterly Review, vol. 13, no. 2, **(2012)**, 25-32.
[5] R. M. Ward, R. Schmieder, G. Highnam, and D. Mittelman, "Big data challenges and opportunities in high-throughput sequencing. Systems Biomedicine", vol. 1, no. 1, **(2013)**.
[6] R. Y. Zhong, Q. Y. Dai, K. Zhou, and X. B. Dai, "Design and Implementation of DMES Based on RFID". Proceeding of 2nd International Conference on Anti-counterfeiting, Security and Identification, Guiyang, **(2008)** August 20-23.
[7] R. Y. Zhong, G. Q. Huang and Q. Y. Dai, "A Big Data Cleansing Approach for n-dimensional RFID-Cuboids". Proceeding of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2014), Taiwan, **(2014)** May 21-23.
[8] A. R. Syed, K. Gillela and C. Venugopal, "The Future Revolution on Big Data". International journal of Advanced Research in Computer and Communication Engineering, vo. 2,no. 6, **(2013)**, pp**.** 2446-2451.
[9] R. Y. Zhong, G. Q. Huang, S. Lan, Q. Dai, T. Zhang, and C. Xu, "A two-level advanced production planning and scheduling model for RFID-enabled ubiquitous manufacturing", Advanced Engineering Informatics, vol. 29, issue 4, **(2015),** pp. 799-812.
[10] S. Sellars, P. Nguyen, W. Chu, X. Gao, K. L. Hsu and S. Sorooshian, "Computational Earth Science: Big Data Transformed Into Insight". Eos, Transactions American Geophysical Union, vol. 94, no. 32, **(2013)**, pp. 277-278.
[11] J. Elder IV and T. Hill, "Practical text mining and statistical analysis for non-structured text data applications," Academic Press. **(2012)**.
[12] X. Qiu, H. Luo, G. Xu, R. Zhong, and G. Q. Huang, "Physical assets and service sharing for IoT-enabled Supply Hub in Industrial Park (SHIP)", International Journal of Production Economics, vol. 159, **(2015)**, pp. 4-15.
[13] J. Staš, D. Hládek, M. Pleva, and J. Juhár, "Slovak language model from internet text data Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces," Theoretical and Practical Issues: Springer. **(2011)**, pp. 340-346.
[14] F. Pasquale. Grand Bargains for Big Data: The Emerging Law of Health Information. Md. L. Rev., 72, 682-1039, **(2013)**.
[15] G. Q. Huang, R. Y. Zhong, and K. L. Tsui, "Special issue on 'Big data for service and manufacturing supply chain management", International Journal of Production Economics. vol. 165, **(2015)**, pp. 172-173.
[16] R. Y. Zhong, G. Q. Huang, Q. Y. Dai, and T. Zhang, "Mining SOTs and Dispatching Rules from RFID-enabled Real-time Shopfloor Production Data", Journal of Intelligent Manufacturing, vol. 25, **(2014)**, pp. 825-843.
[17] R. Y. Zhong, Z. Li, A. L. Y. Pang, Y. Pan, T. Qu, and G. Q. Huang, "RFID-enabled Real-time Advanced Planning and Scheduling Shell for Production Decision-making", International Journal of Computer Integrated Manufacturing, vol. 26, **(2013)**, pp. 649-662,.
[18] R. Y. Zhong and G. Q. Huang, "RFID-enabled Learning Supply Chain: A Smart Pedagogical Environment for TELD, International Journal of Engineering Education", vol. 30, **(2014)**, pp. 471-482,.

Authors

Fan Linxiu, she received bachelor's degree in Computer Science and Technology from Jiangxi Normal University 2003 and master degree in Computer Application from Huazhong University of Science and Technology. She is current a lecturer in Gannan Normal University. Her research interests are computer networks and computer applications such as Big Data Analytics. She has published several papers on journals and conferences.