A Rhombic Dodecahedron Topology for Human-Centric Banking Big Data

Hao Wang, *Member, IEEE*, Shenglan Ma, and Hong-Ning Dai, *Senior Member, IEEE*

***Abstract*— Banks are collecting unprecedentedly large amount of data about their customers from difference sources, considering their cyber, physical, social activities. The focus of this paper is to study the problem of information sharing and lower the communication overhead among different nodes for a specific data mining approach in distributed big data architectures. This problem can be abstracted as how to efficiently search under a specific cluster node topology. This paper proposes a new design rule for topologies including 1) low coordination number, 2) high packing density, and 3) having a 3D structure. According to this rule, a Rhombic Dodecahedron topology is proposed. A distributed banking big data mining framework based on the proposed topology is implemented. The experiments based on multi-optimization benchmark functions show the excellent searching ability of the proposed topology; and a banking customer feature reduction prototype has been implemented to showcase the practicality of the data mining framework.**

***Index Terms*— Financial Big Data; Cyber-Physical-Social Systems; Swarm Optimization; Rhombic Dodecahedron**

I. INTRODUCTION

FINANCIAL organizations like banks are moving quickly towards more human-centric financial services for their customers. Therefore, these organizations are collecting unprecedentedly large amount of data about their customers from difference sources, considering their cyber, physical, social activities. Big data capabilities [1] are gradually becoming the core competitiveness of banks. An important precondition for realizing the value of big data is to be able to reveal the truth and find valuable patterns and insights from these vast amounts of data about their customers [2]. It is difficult to accomplish this by relying solely on the experience and wisdom of experts. It requires a variety of data mining techniques [3]. For example, Citibank established a big data analysis platform for its retail business. This approach has greatly increased the capability to analyze and process data and has significantly influenced Citibank's transforming and upgrading [4]. HSBC uses the data mining tool to find the cross selling and "roll" sales [5].

The technologies that currently have been involved in big data systems [6] include massively parallel processing (MPP) databases [7], data mining [8], distributed file systems [9], distributed databases [10], cloud computing platforms [11][12] and scalable storage systems [13]. Due to massive data volume and various data dimensions, distributed computing platforms are expected to be used [14]. For example, MapReduce [15] is a well-established distributed platform, on top of which data mining algorithms can be effectively executed. However, one of the critical performance bottlenecks lies in optimizing the search procedure in the large scale solution searching space, especially in banking big data framework. Most of current studies mainly concentrate on the performance improvement brought by computing devices.

Therefore, the focus of this paper is to investigate information sharing and the communication overhead reduction among different nodes for a specific data mining approach in distributed big data platforms. The solution to this problem can be redirected to the effective searching under a proper topology of searching particles in swarm optimization [16]. The topological structure of the cluster nodes can be thought as a deep social network. The local neighborhood could affect the behavior of each mining nodes and control the whole cluster's exploration (divergence) versus exploitation (convergence) tendencies. With the relationship of nodes, the searching ability of clusters is essentially affected by the communication capacity of topologies.

In this paper, we propose using *Rhombic Dodecahedron* topology in swarm optimization to improve the searching efficiency in banking big data mining on top of distributed platforms. The contributions of the paper can be summarized as follows:

1) We propose new design rules for selecting topologies with consideration of the following metrics: the coordination number, the packing density, and the 2D/3D structure. We find the good topology should have the low coordination number, high packing density and 3D structure.

2) We propose *Rhombic Dodecahedron* topology in swarm optimization. The Rhombic Dodecahedron topology can fulfill the above design rules. In particular, compared with existing topologies, the proposed Rhombic Dodecahedron has lower

This work is partially funded by the Fujian Fumin Foundation and is partially supported by the National Natural Science Foundation of China under Grant (No. 61672170), the Science and Technology Planning Project of Guangdong Province (No. 2017A050501035), and Science and Technology Program of Guangzhou (No. 201807010058).

Hao Wang is with Department of Computer Science, Norwegian University of Science & Technology, Gjøvik, Norway (e-mail: hawa@ntnu.no).

Shenglan Ma is with Division of Science and Technology, Fujian Rural Credit Union, Fujian, China (e-mail: mashenglan@fjnx.com.cn).

Hong-Ning Dai is with Faculty of Information Technology, Macau University of Science and Technology, Macau (e-mail: hndai@ieee.org).

coordination number, higher packing density and higher chance to reach global optimum due to the usage of 3D structure.

3) We have implemented a prototype of banking big data framework based on the proposed Rhombic Dodecahedron topology, data mining algorithms and distributed computing platforms. In particular, we propose MapReduce Searching Algorithm based on sphere packing topology (essentially based on Rhombic Dodecahedron). Extensive experiment results verify the effectiveness of the framework.

The rest of the paper is structured as follows, Section II presents a typical banking big data mining framework and related work. Section III reviews the concept of bond energy on the topology, presents a new design rule for topologies, and presents a Rhombic Dodecahedron topology with its searching ability evaluated. Section IV uses the proposed topology to design a banking big data mining framework. Section V present 1) the experiments evaluating the searching ability of the proposed topology based on multi-optimization benchmark functions; and 2) a banking customer feature engineering prototype to showcase the practicality of the data mining framework based on the proposed topology. Finally, we draw our conclusions in Section VI.

II. Banking Big Data Framework and Related Work

In this section, we first introduce the typical banking big data framework in Section II-A. We then present related work on banking data mining in Section II-B and topologies in swarm optimization in Section II-C.

A. Banking Big Data Framework

A typical banking big data framework consists of a data access layer, a data exchange layer, a data service layer and a data application layer, as shown in Fig 1. In particular, the data access layer collects internal and external data which has then been submitted to the data exchange layer for the further preprocessing, consequently being saved at the data service layer. It is worth mentioning that the data service layer is composed of MPP databases, transactional databases, the Hadoop platform [21] and Spark platform [22][23] to implement data storage and offer service interfaces. Specifically, the transactional database mainly deals with online business data and adapts to a large number of business scenarios, in which business data requires frequent operations such as *add*, *delete* and *modify*.

The MPP databases serving as a back-end database engine mainly for high-value-density structured data processing can adapt to business scenarios such as batch data processing and data query and analytics. The Hadoop platform is mainly responsible for low-value-density data processing, such as data collection from the Internet. The Hadoop platform can exchange data with transactional databases and MPP databases through high-speed data exchange channels. The Hadoop platform is always used to conduct the distributed data mining tasks.

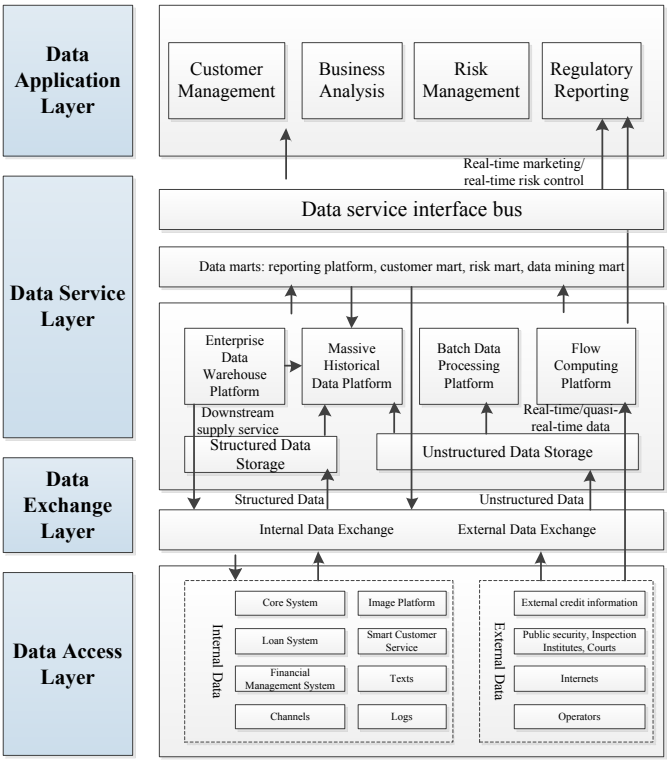

Fig. 1. Banking Big Data Mining Framework

The data service interface bus offers standard service interfaces internally and externally. It includes batch data services, real-time data services, data view services, mobile analysis services, and self-service analysis. The data service interface bus allows the application layer to invoke the data mining application of the data service layer. On the basis of the hybrid architecture, the data service interface provides the enterprise-level data applications for banks and enhances data value. In addition, it also supports the external data application layer via implementing customer management, business analysis, risk management and regulatory reporting.

B. Banking Data Mining

Data mining as an information processing technology is to extract, transform, analyze and model the data in the database to obtain information that is beneficial to decision-making. Banks have many ways to mine big data [24]. The most common data mining methods can be categorized as follows:

1. *Taxonomy*. Banks classify data into different definite categories according to the characteristics of them, and use them to analyze customer classification, customer attributes and customer satisfaction.

2. *Regression analysis method* [25]. This method includes the trend characteristics of data series, the prediction of data series and the correlation between data. According to the regression analysis, the banks can forecast sales trends and develop the targeted promotions by analyzing customer needs and product life cycle.

3. *Clustering analysis* [26]. According to the similarity banking data samples, we can put the data samples in the same

category if they are close to each other. Clustering analysis can be applied to classify customer groups, analyze customer background, predict customer purchase trends and conduct market segmentation.

4. *Association rules* [27]. Through mining corporate customer data, we can identify the existing relationships, analyze the key factors affecting the effectiveness of marketing, and provide product positioning price, risk assessment and fraud prediction in the customer relationship management system.

Data mining methods typically are involved with feature analysis, deviation analysis and swarm intelligence algorithms [28]. In addition, banking data mining is typically based on data warehouse and on-line analysis processing. Banding data mining methods are also running on top of the Hadoop platform or other large-scale data processing platforms (as shown in Fig 1). It uses data mining techniques combined with multiple statistical analysis methods [29] to clean, convert and load, etc. Moreover, Spark is also used for real-time data mining [30][31].

The data processing method discovers the relationships and trends, and completes tasks such as data analysis, knowledge discovery, decision support and financial intelligence. Therefore, the banking data mining execution process has four stages:

1. *Business understanding stage*, in which the needs of the business department should be fully understood, business problems and pain points should be translated into specific business requirements. This stage requires a large number of interviews with business stakeholders, and professional consultants are required to guide business stakeholders to express their own ideas.

2. *Data understanding stage*, in which banking data should be explored to obtain the manifestations of data and the real hidden issues behind business issues with specific business issues should be analyzed.

3. *Data modeling stage*, in which sampling is generally used to divide data set in to the model training set and test set. The training set is used for mining modeling, and the test set is used to test the effectiveness of the model.

4. *Model evaluation stage*, in which the performance of the model should be evaluated based on the three important indicators: accuracy, coverage, and degree of improvement.

C. Topologies of Swarm Optimization

In recent works, many topologies (such as All, Ring, Four clusters, Pyramid, and Square) are discussed [16][17]. The Von Neumann Structure or Square topology is recommended for its good searching ability. However they are just plain lattice and 2D packing, formed by arranging the spheres in a grid, but not close-packing. Its 2D structure makes it hard to find the global optimum when the nodes are searching the local optimum in a different direction. For 'All' topology with all vertexes connected to every other, its coordination number is too big to let the nodes explore the new space, leading to the phenomenon of nodes being easily trapped. This topology is currently used in many Hadoop clusters [18][19]. 'Ring' topology is constructed by connecting every vertex to two others. Its coordination number is too small to exploit the local space efficiently.

Therefore a properly-designed topological structure can make cluster suitably balanced for both exploitation (i.e., convergence) and exploration (i.e., divergence).

III. RHOMBIC DODECAHEDRON

The focus of banking big data mining lies in the searching capability of the distributed computing platforms (e.g., Hadoop). The searching capability essentially comes from the information transmission capabilities of the topology among nodes. To this end, this section will analyze the topology information transmission capabilities from the perspective of crystallology and propose a topological structure suitable for banking big data mining clusters.

A. Design rules of good topologies

This paper proposes the novel design rules inspired from the concept of bond energy in the crystallology. Given the searching space, the information communications among nodes can be characterized by the bond link. In the crystallology, the bond energy is the decisive factor for the exploitation versus exploration tendencies. Although the direct bond energy could not be easily computed in the crystallology, the coordination number and packing density can implicitly reflect the strength of bond energy. The coordination number usually decides the energy to break the bond consequently charging the exploration; packing density affects the exploitation by influencing the searching efficiency in local space. Sometimes, the fitness value imitates the external energy to break the bond link or reconstruct the bond link. Higher fitness can break the link in the crystals and make the search direction redirect to the new space. We give the detailed description on these metrics as follows.

1. Coordination Number

In ionic crystal, the lattice energy is usually used to represent the strength of bonds in these ionic compounds. In other words, the larger lattice energy will make a more stable ionic crystal. The message transferred from one sphere (i.e., the swarm particle) to another is influenced by the ionic bond. The essence of the ionic bond is the electrostatic force between positive and negative ions. If the two ions can be viewed as spheres, then we can conclude that the higher of electric quantity results in the smaller space of two nuclear and the stronger electrostatic interaction will make the stronger ionic bond. Then the ionic crystal turns out to be a more stable structure according to the Coulomb's law:

$$F \propto q_1 \cdot q_2 / r^2$$

where q_1 and q_2 are point charges, and r is the separation distance.

One important factor that reflects this bond energy is the coordination number, which is the number of a central atom's nearest neighbors [32][33][34]. The radius ratio of the ions can affect the coordination number consequently influencing the stability of structures as in [35]. Since $r^+ / r^- > 0.414$, the coordination number is greater than 4, leading to a stable

crystal; while $r^+ / r^- < 0.414$, the coordination number becomes 4, resulting in an unstable crystal. As r^+ continuously grows larger, it can get 12 coordinators; on the contrary if r^+ grows smaller, it only contains 3 coordinators.

2. Sphere Packing and Close-packing Density

In geometry, a *sphere packing* is an arrangement of non-overlapping spheres within a containing space. The considered spheres are usually in the identical size and can have the similar nature *like* nodes in distributed platforms. In particular, the close-packing is a dense arrangement of the equal sphere. Hexagonal close packing and cubic close packing are known to be the densest packing of equal spheres [36]. Every third layer overlying one another arrangement gives the cubic close packing (also called face-centered cubic) and spheres in alternating layers overlying one another gives the hexagonal close packing [37]. The packing density of these two packing arrangements equals to $\pi/(3\sqrt{2}) \approx 0.74048$ since sliding one sheet of spheres cannot affect the volume that they occupy.

3. The New Design Rule

Summing up the above analysis, we then have the novel design rules to guarantee the communication capacity: topologies should have the proper bond energy, such as the low coordination number, the high packing density and 3D structure.

Regarding the coordination number, nodes coordinated with low neighbors have the flexibility to break the corresponding links to search the new space. The coordination number 4 could make the crystal unstable and then may have a good exploration. The smaller coordination number may make the crystal be fragile while the larger of this number yields the inflexibility.

With respect to the packing density, it can be used to represent the probability of finding the global optimum. In a 2D structure, it is hard to explore the global optimum space when the whole topology tends to the upper area and the lower area is blind to it. On the contrary, nodes in a 3D structure still have the chance to explore the global optimum in the same condition. It can break the link of the lower plane to attract the whole topology toward the new space as illustrated in Fig. 2.

Therefore, the proper bond energy (i.e., the low coordination number, the high packing density and the 3D structure) can make topologies have the better communication capacity. It always influences the timestamp to break the bond link in a specified topology.

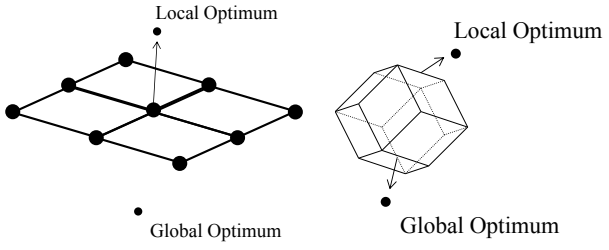

Fig. 2. Nodes of 2D and 3D Structure Flying Toward Local Optimum

B. Rhombic Dodecahedron Topology

We next describe the proposed Rhombic Dodecahedron topology.

1. Features of Rhombic Dodecahedron topology

Using the above design rules, the paper proposes the Rhombic Dodecahedron topology in swarm optimization. In particular, Rhombic Dodecahedron topology consists of twelve congruent rhombuses, 24 edges and 14 vertices. There are two types of vertices: one is made of 4 rhombic acute angles and the other is made of 3 rhombic obtuse angles. The latter is an intersection of 3 rhombuses. The same type of vertexes is impossible to appear on one edge. Fourteen vertices of the Rhombic Dodecahedron are joined by 12 rhombuses as shown in the Fig. 3. The long diagonal of each face is exactly $\sqrt{2}$ times the length of the short diagonal, so that the acute angles on each face measures approximately 70.53°.

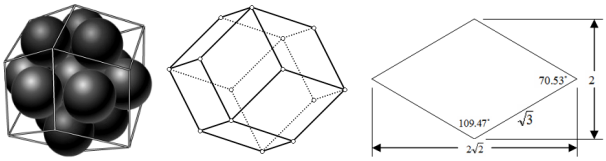

Fig. 3. Rhombic Dodecahedron Topology

This cumulated structure satisfies close packing [38] in the 3D space with 74.05% packing density that makes the full use of space to reach the maximum space utilization. Rhombic dodecahedron can fill the space seamlessly with copies of it, gluing faces together as shown in Fig. 4. And the average coordination number is 3.43 close to 4. Therefore, this topology satisfies the design rules.

2. Comparisons among different topologies

The square topology with Von Neumann structure is solely a plane lattice with 2D packing and is formed by arranging the spheres in a grid. But it is not close-packing in the 2D space since 2D close packing means one sphere surrounded with 6 spheres in the plane (i.e., the hexagonal lattice). Fig. 4 shows the square lattice and the hexagonal lattice

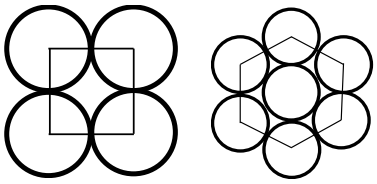

Fig. 4. the square lattice and the hexagonal lattice in 2D

The unit cell of the square lattice contains four 1/4-spheres, then the area is

$$A_{\text{spheres in the unit cell}} = \pi r^2;$$

Meanwhile, the area of the unit cell is

$$A_{\text{unit cell}} = 4r^2;$$

Then, the packing density is

$$\eta = \pi/4 = 0.7854.$$

The area of the unit cell of the hexagonal lattice is

$$A_{\text{unit cell}} = 6[1/2*(2r)(\sqrt{3}r)] = 6\sqrt{3}r^2;$$

The unit cell contains six 1/3-spheres and one midpoint sphere, then the area is

$$A_{\text{sphere in unit cell}} = (6*1/3+1)*\pi r^2;$$

and the packing density is

$$\eta = \pi/2\sqrt{3} = 0.9069.$$

Therefore, in 2D circumstance, the packing density of square topology is lower than Rhombic Dodecahedron. Moreover, square cannot have the 3D view in searching.

We will then discuss the situation that the global optimum is outside topology. The ability to handle this situation can indirectly reflect the design rules as the proper bond energy can also charge the timestamp to break the bond link to explore the new space. If one node of Rhombic Dodecahedron tends to fly to the global area, it will transfer its message to the other three nodes by two runs. Then all nodes can determine whether to break the link or not. Now, the energy to break the link is judged by all these four nodes. After the link breaks, the newly passed message would spread to another 11 nodes by the central sphere. In this way, the whole Rhombic Dodecahedron topology will gradually move to the new space. Notice that Rhombic Dodecahedron has rhombus-searching features and needs 4 decision-makers and 2-run delays to break one atomic link to explore the new space. Hence, the structure is by nature a little sluggish, but makes the local space search effective

In the square topology, each node is exactly equal to the sphere. Then if one node tends to fly toward the global optimum, it should firstly inform four neighbors in one run. Then the neighbors break the link together to let the node explore the new space. The square also needs 5 decision-makers and 1-run delay to break four atomic links to explore the new space. This topology has quick respondence toward new space with enough decision makers while breaking too many links also results in instability.

In summary, the 2D structure limits the communication capacity of square topology. Additionally, for 'All' topology one node needs N (i.e., the number of nodes) decision-makers to break N links to explore the new space, so it makes this topology hard to explore the new space. In the banking big data mining task, the structure has the significant time consumption. Moreover, the 'Ring' topology easily breaks the links for demanding only two decision-makers to break only two links, consequently the weak exploitation arising.

As discussed above, the 3D structure of Rhombic Dodecahedron is superior to other topologies in big data mining framework. To verify this conclusion, this paper compares the Square topologies and Rhombic Dodecahedron topologies in multi-optimization problems. For comparison purpose, we define two-type topologies by the population size: 16-Square、20-Square and 1-Rhombic Dodecahedron are "Single Topologies"; 24-Square and 2-Rhombic Dodecahedron are "Complex Topologies" as shown Fig. 5.

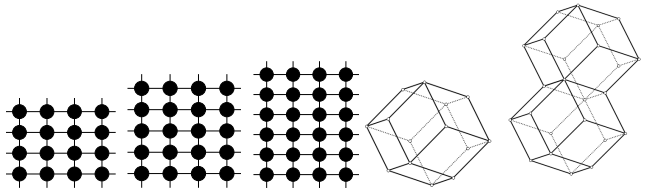

Fig. 5. Square-16, Square-20, Square-24, Rhombic Dodecahedron and 2-Rhombic Dodecahedron

There are three parameters that can be used to evaluate the searching ability of these topologies. Table I lists these three statistics (i.e., the average distance, the diameter, the distribution sequence). The first parameter represents the average number of iterations to broadcast throughout the entire topology, and the second diameter shows the maximum iterations. The third parameter can measure the delay in the information spreading through the topology. Note that the first value of the distribution sequence is the average degree of the graph; it can be regarded as an average coordination number.

TABLE I. GRAPH STATISTICS OF THE TOPOLOGIES

Topology	Average Distance	Diameter	Distribution Sequence
Square-16	2.13	4	<4, 6, 4, 1>
Square-20	2.32	4	<4, 7, 6, 2>
Square-24	2.61	5	<4, 7, 7, 4, 1>
Rhombic Dodecahedron	2.15	4	<3.43, 5.14, 3.43, 1>
2-Rhombic Dodecahedron	2.77	6	<3.67, 6.67, 6.33, 4.33, 1.67, 0.33>

As shown in Table I, we can find that in Rhombic Dodecahedron based topologies, the average number of reachable nodes via directly traversing (i.e., the average coordination number) is lower than the square-based topologies, but the middle traversing process influences more neighbors; this makes Rhombic Dodecahedron a little sluggish while it uses enough decision-making nodes to break the key link to explore the new space. This result also verifies the properties of Rhombic Dodecahedron.

IV. A PROTOTYPE OF BIG DATA MINING FRAMEWORK

This section uses the Rhombic Dodecahedron topology to establish a logical relationship in the Hadoop cluster at the data service layer of the bank big data framework (as shown in Fig. 6, corresponding to the data service layer in Fig. 1). We then propose a data mining algorithm for *MapReduce* components based on the proposed prototype.

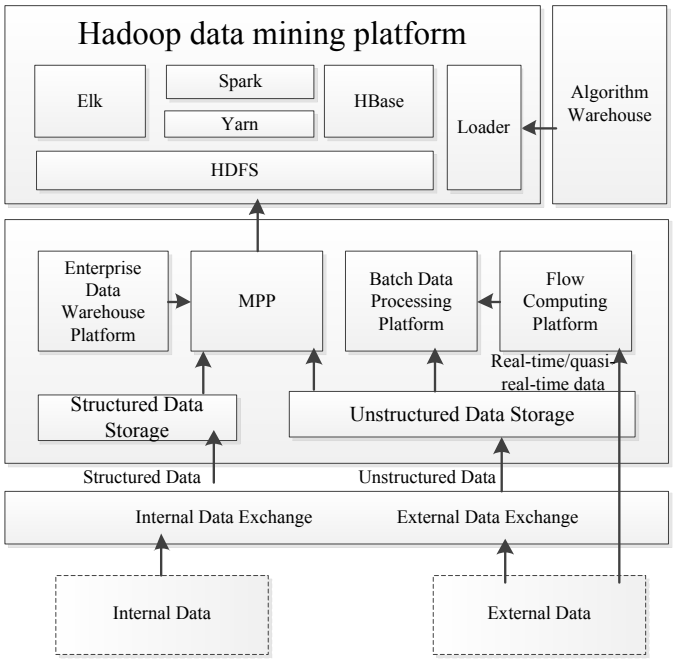

Fig. 6. Hadoop based Banking Big Data Mining Framwork

Fig. 6 shows the Hadoop data mining platform in the big data framework. For the bank's internal structured data, CDC, Flume, and the Sqoop [39] technologies are used to collect customer transaction information from core system log information in real time, including transaction information such as transaction channels, transaction amounts and counterparties. The information cached in Kafka message middleware is then loaded into the big data processing engine and finally stored in the MPP database using the data warehouse cleaning process. The external unstructured data [40] is processed by the Spark stream calculation engine. The customer tag system is built and stored in the MPP database. Hadoop data mining platform extracts data from MPP and conducts mining based on business models to form customer marketing and risk management [41].

Hadoop data mining platform consists of *TaskTracker* and *JobTracker* clusters as shown in Fig. 7.

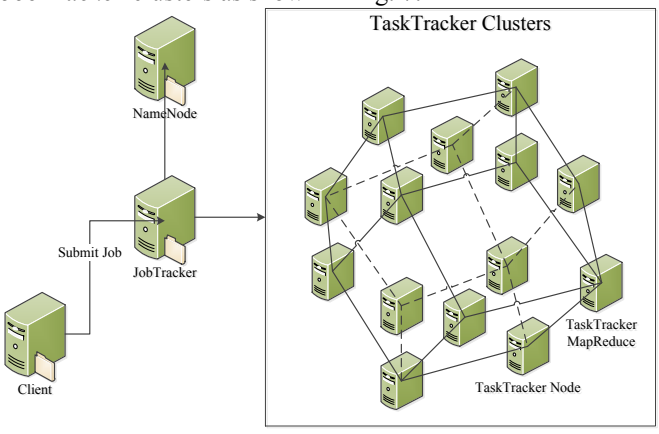

Fig. 7. Hadoop Cluster based on Sphere Packing Topology

TaskTracker is a Hadoop computing process running on the *DataNode* of the Hadoop cluster. The main task of *TaskTracker* is to run the actual computation tasks assigned by the *JobTracker*, such as running *Map* and *Reduce* functions. When the *TaskTracker* receives a task assigned by the *JobTracker*, each map and reduced task is run in a separated JVM process. The *TaskTracker* will send heartbeat messages to the *JobTracker* during the running of the task. The heartbeat message also contains information such as the number of currently free slots. The role of the *JobTracker* process is to run and monitor MapReduce jobs. *NameNode* will initially initialize the logic relationship of the *TaskTracker* nodes, i.e., a Rhombic Dodecahedron topology.

Regarding the solving process of the data mining search problem, the solution of each *Reduce* function is sent to the logically connected *TaskTracker* nodes by the *JobTracker* when each iteration task is completed. Algorithm 1 gives the details on MapReduce Searching procedure based on sphere packing topology.

Algorithm 1: MapReduce Searching Procedure based on Sphere Packing Topology

Input: Job, logic topology
Output: Global Optimal Solutions, Job

1. Client in data application layer call for data mining job. *JobTracker* receives *Job* request;
2. The *JobTracker* requests the list of nodes by *NameNode*, retrieves the sphere packing topology.
3. Initialize each node's local optimal solution data set S_i;
4. The JobTracker determines the execution plan of the Job. It calculates the number of tasks for *Map* and *Reduce* functions that execute the job. According to the logic topology, the $\bigcup_j S_j$ is allocated to the logically connected nodes;
5. JobTracker submits all tasks to each TaskTracker node. The TaskTracker will periodically send a heartbeat to the JobTracker. If the heartbeat is not received within a certain period of time, the JobTracker will consider the TaskTracker node as failed. The JobTracker will then redistribute the task on this node to other nodes to reconstruct the topology;
6. Each *TaskTracker* internally uses $\bigcup_j S_j$ to perform data mining calculations (*Map* function), and then through compute optimal search results R_i of node i (*Reduce* function) among the local topology based on the sphere packing topolgoy;
7. Once all the tasks have been executed, the *JobTracker* will update the status of the job in the current round. If a certain number of tasks fail to execute, the job will be marked as failed;
8. *JobTracker* calculates the optimal solution G of all nodes, if the threshold is satisfied, step 9 is performed, otherwise $S_i = R_i$, step 4 is performed;
9. *JobTracker* sends optimal solution G and runs status of *Job* to Client.

V. EVALUATION

We conduct extensive experiments to evaluate the performance of the proposed approach. In particular, we divide the experiments into two groups: the first group is to evaluate the topologies with optimization problems via Particle Swarm Optimization (PSO); the second group is to evaluate the performance of the proposed sphere packing topology on top of big data framework.

A. Topologies Performances on Some Optimization Problems

1. Three dependent variables comparisons

Three dependent variables were used to test the performance of the specified topologies of the PSO [16]. The

first dependent variable is the standardized performance for the speed of finding the best part of a locally optimal region. The second dependent variable is the median number of iterations required to reach a criterion to indicate the speed. The third dependent measure gives the proportion of successes that meet the criteria within 10,000 iterations. The tested function and five kinds of algorithm types in this part are defined in the paper. Tables II, III and IV show the results of three parameters. In particular, the best results of each algorithm are highlighted in bold. In order to briefly address the results, we mark the algorithm 'Canon' as 'C', 'FIPS' as 'FI', 'wFIPS' as 'wFI', 'wdFIPS' as 'wdFI', 'Self' as 'S', 'wSelf' as 'wS', 'Canonasym' as 'C-asym', 'FIPSasym' as 'FI-asym', and 'wFIPSasym' as 'wFI-asym'.

TABLE II. STANDARDIZED PERFORMANCES OF THE TOPOLOGIES AND ALGORITHMS

	(1)	(2)	(3)	(4)	(5)
C	-0.5574	**-0.5786**	-0.5631	-0.4273	-0.5263
FI	**-0.5112**	-0.5120	-0.4171	-0.4177	-0.4008
wFI	-0.5116	-0.5523	-0.5234	-0.4625	**-0.5712**
wdFI	-0.3856	**-0.7173**	-0.5583	-0.5706	-0.5819
S	-0.4716	-0.4797	-0.3980	**-0.6243**	-0.5984
wS	-0.4166	-0.6010	-0.5727	-0.6849	**-1.0167**
C-asym	-0.5886	-0.4863	-0.5535	**-0.6355**	-0.6147
FI-asym	-0.4558	-0.4843	**-0.6671**	-0.5607	-0.6257
wFI-asym	-0.5016	**-0.5787**	-0.5223	-0.4052	-0.4942

(1)- Square-16
(2)- Square-20
(3)- Square-24
(4)- Rhombic Dodecahedron
(5)- 2-Rhombic Dodecahedron

The single topologies find the fitness peaks quicker than the complex topologies because of small structures. Table II shows that Rhombic Dodecahedron shows very good performance when using *FIPS*、*Self*、*wSelf*、*Canonasym* and *wFIPSasym* among all compared algorithms. This result implies that Rhombic Dodecahedron topology is suitable to get on a fitness peak.

TABLE III. MEDIAN NUMBER OF ITERATIONS TO CRITERIA

	(1)	(2)	(3)	(4)	(5)
C	542.83	**489.33**	567.50	566	515
FI	321.50	**301.00**	368.83	408.33	430.17
wFI	309.67	**281.83**	326.17	379.33	361.67
wdFI	328.33	**305.67**	366.67	419.67	404.33
S	336.50	**307.83**	366.33	429.83	445
wS	**424.67**	461.50	824.33	1317.33	1395.83
C-asym	∞	∞	∞	∞	∞
FI-asym	∞	547.83	**528.00**	627.67	546.67
wFI-asym	∞	461.00	**421.83**	428.00	425.67

(1)- Square-16
(2)- Square-20
(3)- Square-24
(4)- Rhombic Dodecahedron
(5)- 2-Rhombic Dodecahedron

As shown in Table III, we observe that Square-based topologies are rather fast, and Rhombic Dodecahedron based topologies are slower because of having to search one more dimension but not crucial. As analyzed in Section IV, Rhombic Dodecahedron is able to find a good point on a local optimum within a limited time.

TABLE IV. PROPORTION OF EXPERIMENTS REACHING CRITERIA

	(1)	(2)	(3)	(4)	(5)
C	94.17	96.25	95.83	95.42	**97.92**
FI	95.83	99.17	99.17	97.50	**100**
wFI	97.08	97.92	98.75	97.92	**99.68**
wdFI	**99.58**	**99.58**	**99.58**	98.33	**99.58**
S	97.92	97.92	**100**	98.33	99.58
wS	98.75	**100.0**	99.17	97.50	95.67
C-asym	76.67	83.33	**84.17**	76.25	83.75
FI-asym	82.08	92.50	97.92	94.17	**98.33**
wFI-asym	84.17	92.50	96.25	92.08	**98.33**

(1)- Square-16
(2)- Square-20
(3)- Square-24
(4)- Rhombic Dodecahedron
(5)- 2-Rhombic Dodecahedron

Table IV shows that 2-Rhombic Dodecahedron finds the global optimum with the higher proportion than other topologies. In particular, for those difficult asymmetric searching tasks, Rhombic Dodecahedron based model is good at solving them with strong searching ability. Therefore, Rhombic Dodecahedron gets better solutions than Square with only little longer searching time.

2. Detailed comparisons between Square and Rhombic Dodecahedron

We then give more detailed comparisons between these two types of topologies in the fully informed model. Tested functions details could be found in [42]. Table V shows the major details of the tested functions and the Table VI gives the average results in 40 runs in these functions.

TABLE V. BENCHMARK FUNCTIONS, WHERE N IS THE DIMENSION OF THE FUNCTION, F_{MIN} IS THE MINIMUM VALUE OF THE FUNCTION AND $S \subseteq R^n$

No.	Name	N	S	f_{min}
1	Sphere Model	30	$[-100,100]^n$	0
2	Schwefel's Problem 2.22	30	$[-10,10]^n$	0
3	Schwefel's Problem 1.2	30	$[-100,100]^n$	0
4	Schwefel's Problem 2.21	30	$[-100,100]^n$	0
5	Generalized Rosenbrock's Function	30	$[-30,30]^n$	0
6	Step Function	30	$[-100,-100]^n$	0
7	Quartic Function i.e. Noise	30	$[-1.28,1.28]^n$	0
8	Generalized Rastrigin's Function	30	$[-5.12,5.12]^n$	0
9	Ackley's Function	30	$[-32,32]^n$	0
10	Generalized Griewank Function	30	$[-600,600]^n$	0
11	Generalized Penalized Functions	30	$[-50,50]^n$	0
12	Generalized Penalized Functions	30	$[-50,50]^n$	0
13	Shekel's Foxholes Function	2	$[-65.536,65,536]^n$	1
14	Kowalik's Function	4	$[-5,5]^n$	0.0003075
15	Six-Hump Camel-Back Function	2	$[-5,5]^n$	-1.0316285
16	Branin Function	2	$[-5,10] \times [0,15]$	0.398
17	Goldstein-Price Function	2	$[-2,2]^n$	3
18	Hartman's Family	3	$[0,1]^n$	3.86
19	Hartman's Family	6	$[0,1]^n$	-3.32
20	Shekel's Family	4	$[0,10]^n$	-10
21	Shekel's Family	4	$[0,10]^n$	-10
22	Shekel's Family	4	$[0,10]^n$	-10

TABLE VI. RESULTS ON TEST FUNCTIONS

No.	(1)	(2)	(3)	(4)	(5)	(6)
1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	0.019040	0.000000	0.000000	0.000000	0.000000	0.000000
3	0.000026	0.070089	0.006347	0.026043	0.001593	0.003675
4	0.023242	7.185275	3.879214	5.505371	0.686878	0.012820
5	29.08267	35.84351	38.95277	32.71579	34.36853	28.08084
6	38.77500	1.450000	0.675000	0.600000	0.125000	0.050000
7	0.015168	0.035187	0.017541	0.037777	0.009584	0.006362
8	77.33296	23.10791	16.11833	21.73983	12.86362	12.31764
9	4.338334	0.168574	0.000000	0.104322	0.000000	0.000000
10	0.126343	0.004974	0.002402	0.004053	0.001355	0.001296
11	0.462860	0.038874	0.015550	0.015550	0.005182	0.002591
12	0.515549	0.134939	0.003570	0.041584	0.089936	0.000274
13	1.220578	3.236172	2.203886	1.468004	1.022854	0.998003
14	0.000603	0.000713	0.000674	0.000805	0.000810	0.000826
15	-1.03162	-1.03162	-1.03162	-1.031628	-1.03162	-1.031628
16	0.397887	0.397888	0.397887	0.397887	0.397887	0.397887
17	3.000000	3.000000	3.000000	3.000000	3.000000	3.000000
18	-3.86278	-3.86278	-3.86278	-3.862782	-3.86278	-3.862782
19	-3.15700	-3.20985	-3.20697	-3.211554	-3.21057	-3.211272
20	-4.95820	-6.13099	-7.59478	-6.110486	-7.87306	-7.487201
21	-6.48784	-9.27994	-10.0441	-9.968637	-10.2116	-10.23596
22	-6.23700	-9.97293	-10.3335	-10.00982	-10.3336	-10.34477

(1)- Canon
(2)- Square-16
(3)- Square-20
(4)- Rhombic Dodecahedron
(5)- Square-24
(6)- 2-Rhombic Dodecahedron

In Single Topologies, compared with Canon, Square-20 and Square-16, Rhombic Dodecahedron PSO has the lowest average values in function 1, 2, 6, 16, 17, 18 and 19. Compared with Square-16, this topology yields the closer results with minimum results in function 1-6, 8-13, 15-19 and 21-22, only except for Function 7, 14 and 20. In complex topologies, 2-Rhombic Dodecahedron could find better results than Square-24 in most functions. Fig. 8 and Fig. 9 show the graphs of the comparisons based on the average normalized results. In normalized situation, where "Performance" in the vertical axis means the lower value means better performance.

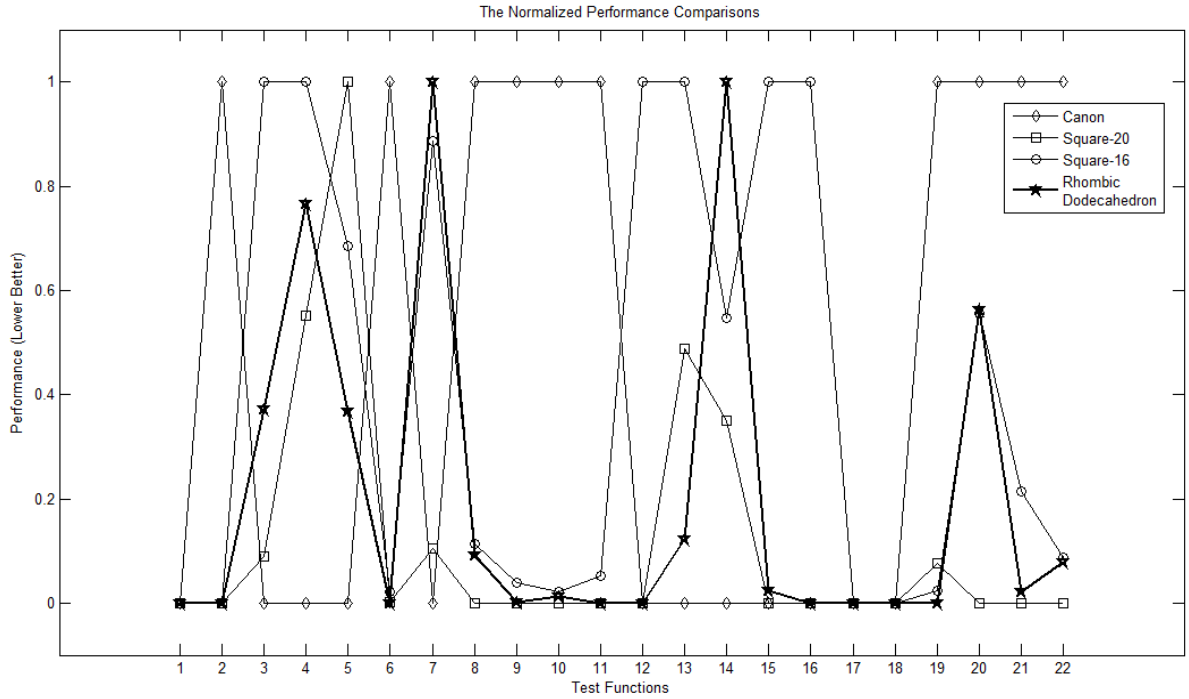

Fig. 8. Performance comparisons of single topologies

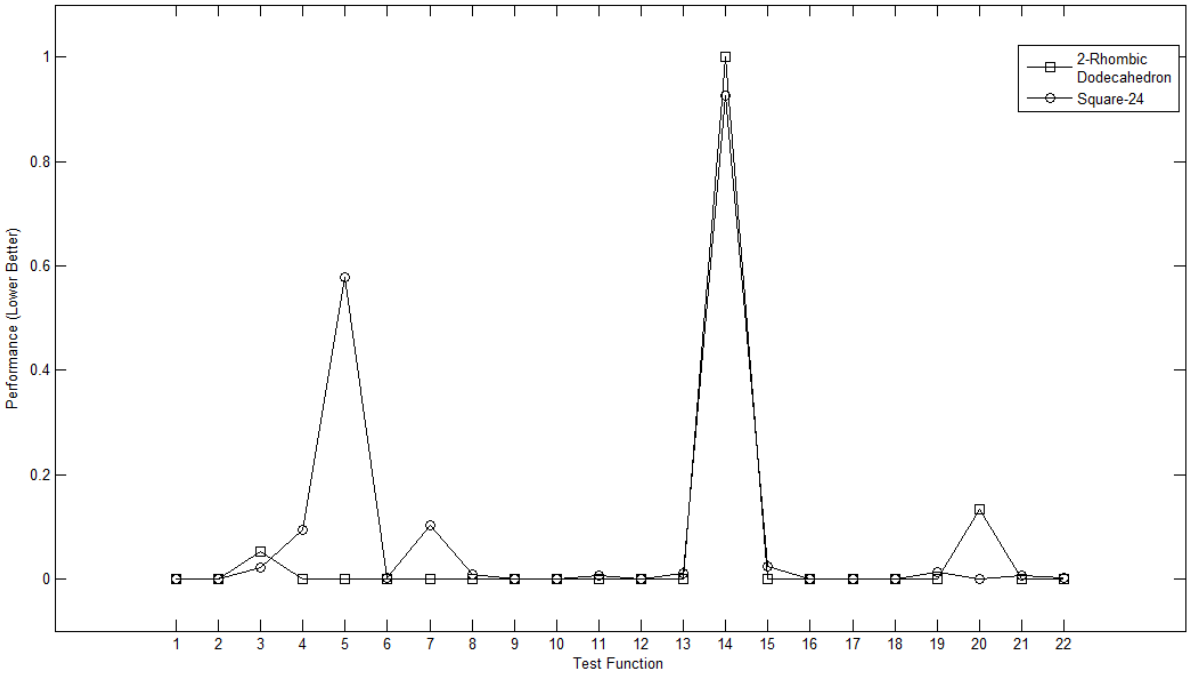

Fig. 9. Performance comparisons of complex topologies

Fig. 8 shows that functions 7 and 14 of the Rhombic Dodecahedron PSO are in the highest level. However, from the raw results, the average results are very close to the minimum results. We also find that Square-20 has the lower line, which shows that the number of nodes influences the searching ability distinctly. However, when compared with Square-16 and Rhombic Dodecahedron only, the latter line is lower than the former line. Although the number of Square-16 is larger, the searching ability is poorer than that of the Rhombic Dodecahedron.

As shown in Fig. 9, 2-Rhombic Dodecahedron is slightly weaker in function 3, 14 and 20. In other functions, it is close to the optimized results and better than Square-24.

Based on the above discussions, we can conclude that the Rhombic Dodecahedron topology has the strong searching ability and yields the good fitness peak. If the number of nodes is not a considerably influencing factor and the task is focused on the ability, Rhombic Dodecahedron topology is preferred. When the number of nodes also matters (in complex topologies), 2-Rhombic Dodecahedron is recommended. In banking big data framework, the nodes of the Hadoop platform are quite huge, so 2-Rhombic can be widely used to improve the performance. Therefore, the proposed topology is suitable for the banking Hadoop cluster.

B. Banking Customer Information Feature Reduction

Bank customer segmentation has far-reaching significance for business marketing. Customer information has the characteristics of large amount of data, high dimensionality, and frequent changing requirements. Therefore, a fast attribute reduction algorithm needs to be introduced to meet the rapid extraction of key attributes and then build it. Using the rough set [43] can maintain the semantic characteristics of the customer data itself, so this section will build a rough set based feature selection search test based on the proposed sphere packing topology on big data mining model.

The testing data set selects customer information in a bank's enterprise customer information factory. The conditional attributes are fetched by customer portrait tags through experience knowledge. This data set consists of 71 attributes such as customer age, gender, education, marital status, industry, position, hobbies and interests, income attributes, living status, car status, aging, activity, loyalty, possession of card products, wage inefficient reserving, large-value idle customers, individual loan risk, post-loan inspection, financial risk, fund risk, etc. Customer value level is selected as the decision attributes. Customers whose total bank income is greater than 0 can be divided into three levels, the top 20% of the revenue are defined as "high-value customers"; the 20% to 80% of revenues are defined as "medium value customers"; the latter 20% of profit ranking are defined as "low value customers".

In the bank test environment, 3,000,000 desensitized customer data samples were selected, PSO was used as a search algorithm [44], and the evaluation function was based on [45]. Besides, the feature selection experimental parameters of the proposed algorithm are defined as:

S: (key, value) (particle index, particle state: including adjacent nodes, position coordinates, velocity, position value, personal optimal position, individual optimal value)

R: (key, value) (optimal particle index, optimal particle state: including adjacent nodes, position coordinates, velocity, position value, personal optimal position, personal optimal value)

The minimum set of reductions calculated through the Hadoop cluster is shown in Table VII.

TABLE VII. THE MINIMUM SET OF REDUCTIONS

No.	Attributes
1	Customer Age
2	Education
3	Industry
4	Position
5	Income
6	Cross-sell Score
7	Financial Term Preferences
8	Debit Card Spending Preferences
9	Credit Card Spending Preferences
10	Loan Potential Customer
11	Credit Card Potential Customer
12	Debit Card Potential Customer
13	Forex Potential Customer
14	Credit Card High Frequency Transactions
15	Loyalty
16	Investment Preferences

From this result, we observe that the attributes of the risk warning class and the customer retention class in the condition attribute are reduced because of the customer value marketing used as a decision attribute. These key attributes of customer can be used as a customer recommendation and other systems [46][47], and effectively solve the problem of excessive calculation cost caused by too many attributes of the marketing system.

VI. CONCLUSIONS

To process the large number of attributes and large amounts of data in banking systems, the distributed strategy for big data mining architectures is necessary. The bond energy, featured by the low coordination number, the high packing density and the 3D structure, is introduced to evaluate the exploration and exploitation of cluster nodes in banking big data framework. We propose novel design rules for topologies in particle optimization. It bases on exploiting the local searching space efficiently and exploring a new space when needed. Based on these rules, this paper presents a Rhombic Dodecahedron topology for cluster nodes to take the exploration and exploitation into account simultaneously. The Rhombic Dodecahedron topology satisfies 3D-close packing structure and has low average coordination number. The experimental results showed that the Rhombic Dodecahedron topology has better performance in finding fitness peak and global optimum. A complete prototype of big data mining framework of Rhombic Dodecahedron topology is implemented with a detailed MapReduce searching procedure. Finally, a feature reduction search experiment based on big data mining framework is tested, and the computed minimum reduct proves the practicality of the framework.

REFERENCES

[1] Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." International Journal of Information Management 35.2 (2015): 137-144.

[2] Wu, Xindong, et al. "Data mining with big data." IEEE transactions on knowledge and data engineering 26.1 (2014): 97-107.

[3] Kiron, David. "Lessons from becoming a data-driven organization." MIT Sloan Management Review 58.2 (2017).

[4] Chen, Zhuming, et al. "The transition from traditional banking to mobile internet finance: an organizational innovation perspective-a comparative study of Citibank and ICBC." Financial Innovation 3.1 (2017): 12.

[5] XI, Yu-ping, and C. H. E. N. Min. "Application of Data Mining Technology in CRM System of Commercial Banks." DEStech Transactions on Engineering and Technology Research eeta (2017).

[6] Sun, N., et al. "iCARE: A framework for big data-based banking customer analytics." IBM Journal of Research and Development 58.5/6 (2014): 4-1.

[7] Patel, Aditya B., Manashvi Birla, and Ushma Nair. "Addressing big data problem using Hadoop and Map Reduce." Engineering (NUiCONE), 2012 Nirma University International Conference on. IEEE, 2012.

[8] Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[9] Huang, Dan, et al. "DFS-container: achieving containerized block I/O for distributed file systems." Proceedings of the 2017 Symposium on Cloud Computing. ACM, 2017.

[10] Elmasri, Ramez. Fundamentals of database systems. Pearson Education India, 2008.

[11] Hashem, Ibrahim Abaker Targio, et al. "The rise of “big data” on cloud computing: Review and open research issues." Information Systems 47 (2015): 98-115.

[12] Huang, Chengqiang, et al. "Time series anomaly detection for trustworthy services in cloud computing systems." IEEE Transactions on Big Data (2017).

[13] Jaffe, Elliot. Scalable Storage Systems. Hebrew University, 2010.

[14] Angiulli, Fabrizio, et al. "Distributed strategies for mining outliers in large data sets." IEEE transactions on knowledge and data engineering 25.7 (2013): 1520-1532.

[15] Zikopoulos, Paul, and Chris Eaton. Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media, 2011.

[16] Mendes, Rui, James Kennedy, and José Neves. "The fully informed particle swarm: simpler, maybe better." IEEE transactions on evolutionary computation 8.3 (2004): 204-210.

[17] Kennedy, James, and Rui Mendes. "Neighborhood topologies in fully informed and best-of-neighborhood particle swarms." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 36.4 (2006): 515-519.

[18] Huang, Wenzhun, et al. "A novel cluster computing technique based on signal clustering and analytic hierarchy model using hadoop." Cluster Computing (2017): 1-8.

[19] Kovacheva, Zlatinka, et al. "Big data mining: In-database Oracle data mining over hadoop." AIP Conference Proceedings. Vol. 1863. No. 1. AIP Publishing, 2017.

[20] Williams, Robert. The geometrical foundation of natural structure. New York: Dover, 1979.

[21] Mehta, Deepak. "Implementation of Improved Apriori Algorithm on Large Dataset using Hadoop." Asian Journal of Computer Science Engineering (AJCSE) 2.06 (2017).

[22] Zaharia, Matei, et al. "Apache spark: a unified engine for big data processing." Communications of the ACM 59.11 (2016): 56-65.

[23] Spark A. Apache Spark: Lightning-fast cluster computing. URL http://spark. apache. org, 2016.

[24] Su, Fei, et al. "A Survey on Big Data Analytics Technologies." International Conference on 5G for Future Wireless Networks. Springer, Cham, 2017.

[25] Draper, Norman R., and Harry Smith. Applied regression analysis. Vol. 326. John Wiley & Sons, 2014.

[26] Maechler, Martin, et al. "Cluster: cluster analysis basics and extensions." R package version 1.2 (2012): 56.

[27] Adamo, Jean-Marc. Data mining for association rules and sequential patterns: sequential and parallel algorithms. Springer Science & Business Media, 2012.

[28] Zhang, Zhongshan, et al. "On swarm intelligence inspired self-organized networking: its bionic mechanisms, designing principles and optimization approaches." IEEE Communications Surveys & Tutorials 16.1 (2014): 513-537.

[29] Plattner, Hasso. "A common database approach for OLTP and OLAP using an in-memory column database." Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. ACM, 2009.

[30] Bifet, Albert, et al. "StreamDM: Advanced data mining in Spark streaming." Data Mining Workshop (ICDMW), 2015 IEEE International Conference on. IEEE, 2015.

[31] Wu, Yulei, Fei Hu, Geyong Min, and Albert Y. Zomaya, eds. Big data and computational intelligence in networking. CRC Press, 2017.

[32] De, Anil Kumar. A text book of inorganic chemistry. New Age International, 2007.

[33] Hermann, Andreas, Matthias Lein, and Peter Schwerdtfeger. "The search for the species with the highest coordination number." Angewandte Chemie International Edition 46.14 (2007): 2444-2447.

[34] McNaught, Alan D., and Alan D. McNaught. Compendium of chemical terminology. Vol. 1669. Oxford: Blackwell Science, 1997.

[35] Pauling, Linus. "The principles determining the structure of complex ionic crystals." Journal of the american chemical society 51.4 (1929): 1010-1026.

[36] Hales, Thomas C. "A proof of the Kepler conjecture." Annals of mathematics (2005): 1065-1185.

[37] Weisstein, Eric W. "Hexagonal Close Packing." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/HexagonalClosePacking.html

[38] Krishna P, Pandey D, Taylor C A. Close-packed structures. International Union of Crystallography, 1981.

[39] Singh, Vikash Kumar, et al. "A Literature Review on Hadoop Ecosystem and Various Techniques of Big Data Optimization." Advances in Data and Information Sciences. Springer, Singapore, 2018. 231-240.

[40] Blumberg, Robert, and Shaku Atre. "The problem with unstructured data." Dm Review 13.42-49 (2003): 62.

[41] Shvachko, Konstantin, et al. "The hadoop distributed file system." Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on. Ieee, 2010.

[42] Yao, Xin, Yong Liu, and Guangming Lin. "Evolutionary programming made faster." IEEE Transactions on Evolutionary computation 3.2 (1999): 82-102.

[43] Jensen, Richard, and Qiang Shen. "Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches." IEEE Transactions on knowledge and data engineering 16.12 (2004): 1457-1471.

[44] Ma, Shenglan, and Dongyi Ye. "Research on Computing Minimum Entropy Based Attribute Reduction via Stochastic Optimization Algorithms." Pattern Recognition and Artificial Intelligence 1 (2012).

[45] Ye, Dongyi, Zhaojiong Chen, and Shenglan Ma. "A novel and better fitness evaluation for rough set based minimum attribute reduction problem." Information sciences 222 (2013): 413-423

[46] Bahnsen, Alejandro Correa, et al. "Feature engineering strategies for credit card fraud detection." Expert Systems with Applications 51 (2016): 134-142.

[47] Wang, Xiaokang, Wei Wang, Laurence T. Yang, S. Liao, D. Yin and M. Jamal Deen, “A Distributed HOSVD Method with Its Incremental Computation for Big Data in Cyber-Physical-Social Systems”, IEEE Transactions on Computational Social Systems, Vol.5, no. 2, pp. 481-492, 2018.

[48] Wang, Xiaokang, Laurence T. Yang, Huazhong Liu, and M. Jamal Deen, “A Big Data-as-a-Service Framework: State-of-the-Art and Perspectives,” IEEE Transactions on Big Data, Vol. 4, No. 3, pp. 325-340, 2018.

Hao Wang is an associate professor in the Department of Computer Science in Norwegian University of Science & Technology, Norway. He has a Ph.D. degree and a B.Eng. degree, both in computer science and engineering. His research interests include big data analytics, industrial internet of things, high performance computing, safety-critical systems, and communication security. He has published 100+ papers in reputable international journals and conferences. He served as a TPC co-chair for IEEE DataCom 2015, IEEE CIT 2017 and ES 2017, a Senior TPC member for CIKM 2019, and reviewers for journals such as IEEE TKDE, TII, TBD, TCSS, TETC, T-IFS, IoTJ, and ACM TOMM, TIST. He is a member of IEEE IES Technical Committee on Industrial Informatics.

Shenglan Ma received the Master degree in computer science from Fuzhou University, Fujian, China, in 2012. After that, he worked in the Fujian Rural Credit Union (FRCU). He is currently a software requirement manager in FRCU. He was a visiting scholar of NTNU in 2018. His research interests mainly focus on big data and data mining.

Hong-Ning Dai is currently with Faculty of Information Technology at Macau University of Science and Technology as an associate professor. He obtained the Ph.D. degree in Computer Science and Engineering from Department of Computer Science and Engineering at the Chinese University of Hong Kong. His current research interests include big data analytics and internet of things. He has published more than 80 peer-reviewed papers in refereed journals and conferences. He is also a holder of 1 U.S. patent and 1 Australia innovation patent. He is the winner of Bank of China (BOC) Excellent Research Award of Macau University of Science and Technology in 2015. He holds visiting positions at the Hong Kong University of Science and Technology, the University of New South Wales, Sun Yat-sen University, University of Electronic Science and Technology of China and Hong Kong Applied Science and Technology Research Institute, respectively. He has served as a guest editor for IEEE Transactions on Industrial Informatics and an editor for International Journal of Wireless and Mobile Communication for Industrial Systems. He is a senior member of the Institute of Electrical and Electronics Engineers (IEEE) and a professional member of the Association for Computing Machinery (ACM)