International Journal of Database Theory and Application
Vol.10, No.3 (2017), pp.1-10
http://dx.doi.org/10.14257/ijdta.2017.10.3.01

# Current Situation and Application of Graph Data Mining Technology

Mengke Zhang[1],Pingping Wei[2*] ,Suzhi Zhang[2]and Jiaxing Xu[2]

[1]*International School,Beijng University of Posts and Telecommunications,Beijing 100876,China.*
[2]*School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China*
[1]*734247021@qq.com,*[2]*634654012@qq.com*

## *Abstract*

*As an important data structure, graph can be used to describe the complex relationship among stuffs. With the setting up of social network, web network and other network in figure data, data mining technology has gradually become a hot research. Traditional data mining technology has been applied to the field of graph data mining constantly. Consequently the development of the graph data mining technology has been accelerated. This paper demonstrates the definition of graph data, and the current graph data mining algorithms which include graph classification, graph clustering, query graph, graph matching, graph of frequent subgraph mining, and graphic database development status. At last, what challenges graph mining technology confronts is illustrated in this paper.*

***Keywords****: graph,graph data, graph mining.*

## 1.Background Meaning

As an important data structure, graph can refer to the complex relationship among stuffs [1].Various network systems are all in the graph structure formation such as social networks, the worldwide web, academic cooperation network, communication network data and so on.With the development of information technology, the data in these networks is increasing constantly. How to mine the potential information and get constructive value in these data is a problem that needs to be solved urgently. Graph data mining technology as a research hotspot have been widely studied and reviewed by scholars in recent years. Graph data mining technology has the nature of the traditional data mining technology,in addition,it also has the complex relationship between the data objects, abundant data forms,therefore complex data structure is a excellent tool for processing. Using graph data mining technology to mine the data in the system, get the hypothetical information, and apply it to pattern recognition, electronic commerce, finance and other fields.

## 2. Graph Data Definition

The graph is composed of the edges of the points and the connecting points. The connecting points are usually used to represent the user's vertices and the edges of the graph represent the relationship between the users. Graphs include directed graphs and undirected graphs. Edge is generally expressed by two vertices, if the two vertices disorder, it is named undirected graph,As shown in Figure 1 (a); if the two vertices in order, it is named directed graph,As shown in Figure 1 (b). The graph is represented as an ordered two dimensional group $G(V, E)$ , among them:

Project supported by the National Natural Science Foundation of China (Grant No. 61201447)
Corresponding author.
*Email addresses:634654012@qq.com (WeiPingping).

ISSN: 2005-4270 IJDTA

$V$ representation of nodes in a graph: $V=\{v_1,v_2,v_3,\cdots v_n\}$; $E$ represents a collection of edges in a graph: $E=\left\{\left(v_i,v_j\right)\middle|\left(v_i,v_j\right)\in R,1\le i,j\le n,i\ne j\right\}$;

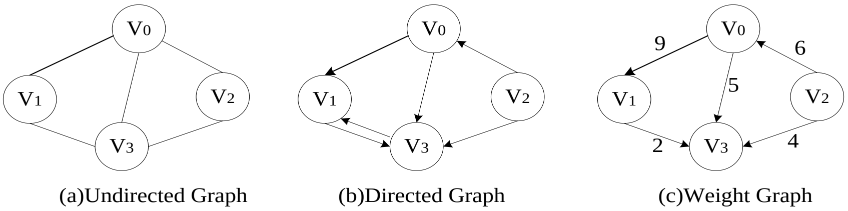
The number of edges V of the vertices is called d(v),the degree of the vertices. If the edge is given to the data information, the weight of the edge is called the intensity of the relationship between the vertices. If the vertex is granted the information, called the attribute of the vertex. This graph is called the attribute weighted graph [2],Weighted attribute graph is widely used in graph clustering.As shown in Figure 1 (c).

Now $G=\left(V,E,A,\omega\right)$;Among them: $V$ represents a node collection in an attribute graph: $V=\{v_1,v_2,v_3,\cdots v_n\}$; $E$ represents a collection of edges in an attributegraph:

$E=\left\{\left(v_i,v_j\right)\middle|\left(v_i,v_j\right)\in R,1\le i,j\le n,i\ne j\right\}$;

$A$ represents a collection of properties of a graph: $A=\left\{a_1,a_2,a_{3,}\cdots a_m\right\}$ ;among them: $\omega$ stands for the weight value of the edge: $\omega=\left\{\omega_1,\omega_2,\omega_3\cdots\omega_p\right\},\left|\omega\right|=p'$.

If K represents the nonempty set of V(G),v represents any vertex, and $\mathrm{M(v)}=\{u\mid u\in K,uv\in E(G)\}$ stands for the field of K,for any graph G or N, if $\mathrm{V}(N)\subseteq V(G)$ and $\mathrm{E}(N)\subseteq E(G)$, then N is called subgraph of G.As Figure 1 illustrates,the three kind of graphs are shown below:



(a)Undirected Graph (b)Directed Graph (c)Weight Graph

**Figure 1. Comparison Graph**

## 3.Research Status of Graph Data Mining Technology

With the deepening of the research of graph data, graph data mining technology has been widely studied and developed. Graph data mining technology mainly include graph classification, graph clustering, graph query, graph matching, graph mining, graph database, etc. This paper mainly introduces the research status and characteristics of these technologies, meanwhile, compares their advantages and disadvantages.

### 3.1 Graph Classification

Graph classification is an important part of graph mining technology which is classified by the classification model. According to whether node labels and training tuple class exist, classification graph will be subdivided into unsupervised classification, supervised classification and semi supervised classification. Because of the difference of the classification model, the method of graph classification includes the classification based on frequent subgraph model, based on the probabilistic substructure model and based on graph kernel function model.

As the classification feature, classification based on frequent subgraph model is to classify the structure and attribute of the frequent subgraph. It mainly consists of three steps: to begin with, we dig out the frequent subgraph; nextly, the feature of frequent subgraph is used as the classification feature; in the end, structural classification model is constructed. The advantage of the algorithm is that it can be adapted to all kinds of graph data, and the structure characteristic of the graph is classified, and the algorithm is simple. The disadvantage is frequent subgraph features and scale is not easy to determine when the characteristics of frequent subgraph size for a little while, will produce a large amount of classification model, classification results lower, lengthen the time of classification and low classification accuracy.

Classification is based on probabilistic substructure model: This method mainly consists of two algorithms: Apriori algorithm [3] and FP-Growth algorithm [4]. The core idea is to classify a complete subgraph model according to the support measure. The advantage of these algorithms is that the algorithm has low time complexity and high accuracy, and the disadvantage is that the efficiency is very low for large input databases.

Classification based on graph kernel model: "Core" relates to the framework of the atomic function associated with the structure of the graph in the graph. Classification based on graph kernel model, is to assign each node in the graph a "case", In the operation of the edge of the graph structure, the calculation of the similarity of the node "case". Through experiments, it is indicated that the similarity of the feature "case" is better than the distance based computing. The current graph kernel classification algorithm consists of a loop and a based on the walk of the graph kernel algorithm. Xiong[5] et al.proposed protein classification based on graph kernel. The shortcomings of the graph kernel algorithm are having limitations, and frequent subgraph classification algorithm as implemented by all of the graph structure, but in the classification effect of specific graph structure is better in frequent subgraph classification algorithm.

### 3.2 Graph Clustering

Graph clustering is to divide the nodes into a cluster [6] in the condition of considering the structure of the edge. After dividing the clusters can better extract and analyze the object to be studied. Graph clustering algorithm is not restricted to the division of the parallel structure. At present, the core research of graph clustering algorithm is based on the structure and properties of the division, in order to achieve better clustering effect. According to the different clustering algorithms, the clustering algorithm is divided into cluster adaptive algorithm and the algorithm based on vertex similarity. Based on the similarity of the vertices include the algorithm is based on the adjacency matrix algorithm, the distance similarity algorithm and the connectivity algorithm. The clustering algorithm includes the algorithm based on the cutting algorithm and the algorithm based on the density. Graph clustering algorithm can also be separated into global clustering algorithm and local clustering algorithm. Global clustering algorithm is to separate all the nodes in the graph into clusters; Local clustering algorithm can only divide a part of the vertices of each cluster. Based on different criteria, graph clustering can be separated into clustering based on vertex structure similarity, clustering based on attribute similarity and clustering based on attribute similarity. At present, the classical graph clustering algorithm is Kernighan-Lin algorithm [7], spectral clustering [8], GN[9], etc.

Kernighan-Lin algorithm is a graph clustering algorithm based on a greedy algorithm.It based on the greedy algorithm complex network is divided into two communities, and join the gain function $p$, divided two community all edges minus the number of the two agencies interval number of edges is $p$, followed by a constant search another division makes $p$ to a maximum value, you can complete the clustering objective. However, the algorithm needs to know the size of the two community, otherwise it cannot get a good clustering effect, the flexibility is poor.

Spectral clustering belongs to point to cluster. The core theory of spectral clustering algorithm is graph theory. The essence of it is to divide the clustering into graph. In the spectral clustering algorithm, the clustering object is the node of the undirected weighted graph, and the similarity of the object features is represented by the weight of the edge. Secondly, the distance matrix of the graph is obtained, and the object coordinate is calculated according to the distance matrix. Finally,clustering based on coordinates. Spectral clustering is an efficient clustering algorithm, but the research of spectral clustering is still in the initial stage, and has not formed a perfect theoretical system, so further research is necessary.

GN algorithm is based on breadth first search algorithm, the shortest path to a node and the other nodes is obtained by a breadth first search, and the number of edges is equal to the number of the shortest path. Dielectric numerical value is greater, the bibliography shortest path, the edges of the greater probability between the two communities, so division basis is constantly shifting ex large boundary value to achieve clustering. And it can be utilized to distinguish between the position of the community. The GN algorithm is applied to the larger network community clustering, but it needs to know the number of the network community before clustering. As Figure 2 illustrates,this graph clustering model is as follows:

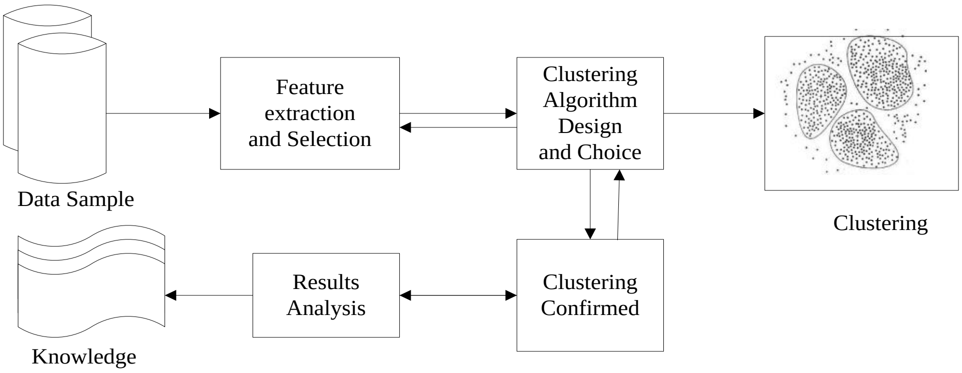

Feature extraction and Selection
Clustering Algorithm Design and Choice
Data Sample
Clustering
Results Analysis
Clustering Confirmed
Knowledge

**Figure2.Clustering process**

### 3.3. Graph Query

Graph query refers to the pattern of the same or similar pattern as the input graph and the input graph in the graph database. Graph query mainly is composed of three aspects: reachability query, distance query and keyword query. Reachability query can be used to determine whether there is a path between nodes; Distance query can be used to obtain the shortest path between nodes; Keyword query is used to study and discover the relationship between the nodes and the superior group of keywords. At present, there are three main problems: how to improve the efficiency of mining graph structure, how to find the subgraph of all the keywords and improve the accuracy of the query. The classical algorithm of graph query is BANKS algorithm and bidirectional query algorithm,however,this kind of algorithm can not know the whole structure of the graph, and the distribution of keywords in the graph can not be obtained because the querying is blind. There is likewise a kind of algorithm is based on the index of graph query algorithm. The representation algorithm is proposed by X.Yan[10]with frequent subgraph as the index structure for graph mining.

### 3.4. Graph Matching

Graph matching is all the subgraphs that match the given input patterns from the data graph. Graph matching is the process of comparing the similarity between graph structure. According to the matching accuracy rate, the graph matching is split into exact graph matching and non exact graph matching. The exact matching method includes the maximum common subgraph, the minimum common subgraph and the subgraph isomorphism. The principal method of non literal matching is to edit distance algorithm.

This paper mainly introduces the method of isomorphism of subgraph and edit distance method.

Sub graph isomorphism is a type of exact matching, given a data graph and an input graph, and only if there exists a subgraph isomorphism with the input graph, the data graph and the input graph isomorphism. Sub graph isomorphism belongs to NP- complete problem. Graph matching efficiency is low. And the subgraph isomorphism requires that the subgraph and the input graph have the exact same graph topology, which reduces the matching range of the subgraph isomorphism.At present, the research on the isomorphism of the subgraph is mainly about how to reduce the constraint points to improve the matching efficiency. Commonly used methods are approximate graph matching, graphic simulation and simulation, etc.

Edit distance is based on the string matching algorithm, which is a method to measure the difference between the graph. Through a series of editing operations on the structural difference between different models, it shows that the difference between different graph structures can be converted into each other by some editing operations. Edit operations include insertion, substitution and deletion of nodes and edges. Edit distance seems to have a lot of classical algorithms, Myers[11] et al. proposed an algorithm about Bayesian based edit distance, Liu[12]proposed the graph matching algorithm which based on edit distance graph.All these algorithms have achieved excellent results.

### 3.5 Frequent Sub Graph Mining

Frequent subgraph mining is a common substructure which is more than the minimum support in the mining graph. Frequent subgraph mining algorithm includes the greedy search algorithm, based on depth first traversal algorithm, based on breadth first traversal algorithm and the maximum frequent subgraph mining algorithm based on the processing of large scale graph.

Greedy search algorithm is based on the minimum description length of frequent subgraph mining algorithm. Generic algorithms are SUBDUE[13], etc.. The core of SUBDUE is the minimum description length, according to the greedy algorithm, using the vertex pattern mining can effectively compress the input data model. SUBDUE supports the discovery of approximate substructures. SUBDUE can be flexibly implemented in the fields such as the definition of social network graph structure,etc..

Most of the breadth first traversal algorithm are based on the frequent subgraph mining algorithm based on Apriori[14], mainly includes AGM(Apriori-based Graph Mining)、FSG(Frequent Subgraph Discovery) [15] ,etc.. In each step, the AGM algorithm increases the size of the subgraph, so as to mine the frequent subgraph. The algorithm is based on the assumption of mathematical recursion, which is suitable for dense graph data. FSG is an improved algorithm of AGM, and the size of the edge is added to the algorithm, and the candidate sub graph is calculated by the optimization method, and the mining efficiency is improved, But the FSG algorithm is only limited to connected graph, which has some limitations.

Depth first traversal algorithm is a frequent subgraph mining algorithm based on pattern growth.Mainly includes gSpan、CloseGraph、FFSM(Fast Frequent Subgarph Mining) and so on. Yan[16] et al. first proposed the gSpan algorithm, the algorithm of the node set of the graph increases, in order to establish a depth first search tree, reduce the generation of copy. CloseGraph algorithm cannot only improve the efficiency of data mining, but also reduce the unnecessary generating subgraph. The efficiency of FFSM algorithm is more important than of gSpan algorithm. The algorithm can deal with the basic problems of subgraph isomorphism, and improves the efficiency of mining.

The maximal frequent subgraph mining algorithm is the algorithm of data mining. These algorithms can improve the effectiveness of data mining, and reduce the scale of the process of mining, Common algorithms are Spin、MARGIN[17]and MFME. The storage space occupied by MARGIN algorithm is higher than Spin algorithm, but the processing efficiency is higher. MFME algorithm to map the edge of the table, in an allusion to, the edge of the table for mining in order to improve the efficiency of mining.

## 4.Graph Database

The database model contains the hierarchical model, the graph model and the relational model. With the arrival of the era of big data, the development of social networks. Data size is increasing. The complexity of data continues to increase. The traditional relational database has been unable to meet the needs of data mining. Graphics database, as a non - traditional relational database, can quickly update data and the relationship between data. The graphics database can perform complex operations efficiently. Therefore, graph database has been developed rapidly and applied.

Graph model is another kind of advance of a hierarchical model. In figure database, the data is stored in a chart. The increase, delete, modify, query and other operations of the traditional database to figure data mining. And social networks, e-commerce and other large amounts of data generated, more suitable for the use of a graph database for storage and operation. Generic graph database is Neo4j[18]. As Figure 3 illustrates,this Neo4j model is as follows:

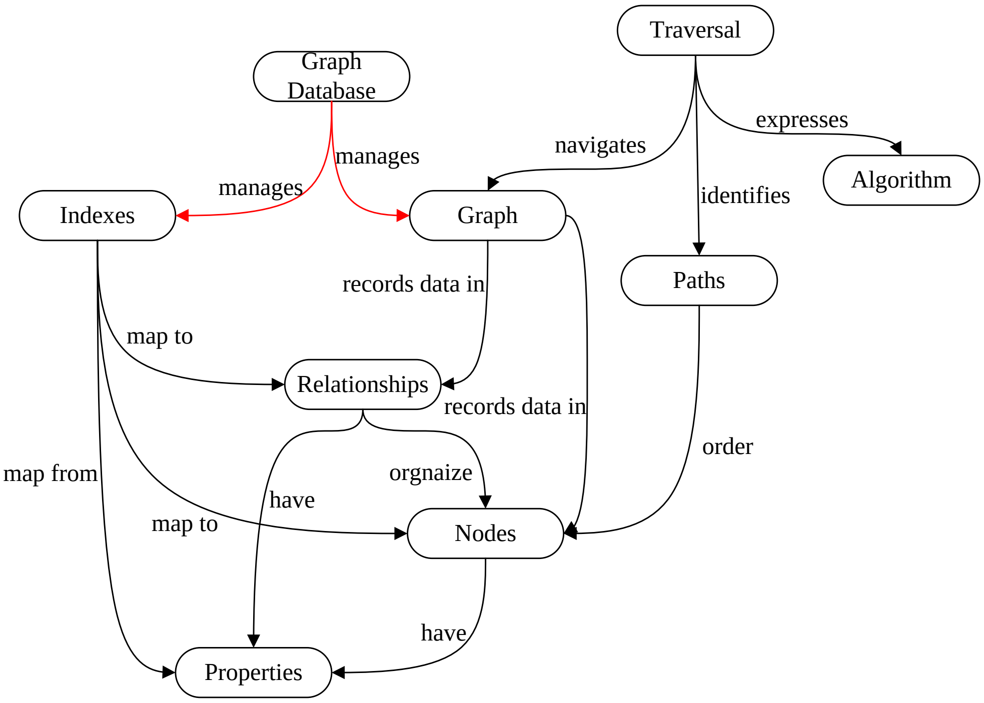


**Figure 3. Neo4j**

Neo4j is an open source database, which is implemented by java language. Neo4j is tantamount to an embedded, and fully real, based on the disk's persistent engine. In neo4j database, data in an attempt to store in the form, the operation of the database more flexible and efficient, especially in the attribute graph processing with high efficiency. Neo4j database is fully consistent with ACID features, compatible with diverse operating systems. When dealing with large-scale social network data, it has the characteristics of low latency, high efficiency and scalability.

## 5. The Challenge of Graph Data Mining

With the research of graph data, the research of graph data mining has made enormous strides. At present, the mining algorithm based on graph clustering and graph classification is becoming more and more mature. Graph search, graph database, graphic modeling, chemical map data and application in bioinformatics will be the hot research in the future. How to implement data mining to the analysis of complex networks is also the research direction in the future. At the same time, graph data mining is facing many challenges:

(1)Scalable graph mining: Graph data mining technology can be used in the memory of the smaller scale of the map data, for a high degree of scalability still has a huge challenge. Therefore, it is important to study the graph mining algorithm based on disk, or a graph mining algorithm based on some parallel processing model, such as DNA model[19], MapReduce[20], etc..

(2)Graph data stream mining: With the development of social network, a large number of data with sudden and graph structural relationships between users at different time points appeared. Data is not stored in the disk, but in the form of data flow structure. How to mine huge scale graph data stream is a challenging problem in the future.

(3)Data mining of uncertain graph：In the process of graph data mining, there is some uncertainty about the relationship of some graph data. How to explore the potential relationship and information of these uncertain graph data is a difficult and challenging problem in graph data mining. There are a lot of theoretical research on the uncertain data mining, which can be applied to the graph data mining.

(4)Mining multiple and heterogeneous graphs: Graph mining is only limited to a single object, how to carry on the multigraph mining is a hot research topic in the future. For example, mining multi between query and single graph with multiple graph structure of the graph. At the same time, it is likewise a fundamental challenge to the heterogeneous graph mining with unique fixed point and edge structure.

## 6. Concluding Remarks

With the development of technology Web2.0, social networks and web network data continue to increase, graph data mining has become a new research hotspot in the data mining area. This paper presents the definition and classification of graph data mining. This paper summarizes the research status of graph classification, graph clustering, graph query, graph matching, graph of frequent subgraph mining and graph database.And analyzes the problems and challenges of graph data mining. Although the graph data mining has produced some valuable research, but figure data mining technology still requires a lot of researchers ask a lot of effort, hope that this paper can play a reference role for the research.

## Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant No.61201447

## References

[1] Ding Y, Zhang Y, Li Z H,” Research and development of graph data mining technology”,Computer Application.vol.32,no.1, **(2012)** ,pp.182-190.

[2] Zhang Nan. “Research and application of weighted attribute graph clustering algorithm based on DNA computing”, East China University of Science and Technology, **(2015)**.

[3] N Li,L Zeng,Q He,Z Shi, "Parallel Implementation of Apriori Algorithm Based on MapReduce", Acis International Conference on Software Engineering, Artificial Intelligence, NETWORKING and Parallel & Distributed Computing IEEE, **(2012)**,pp.236-241.

[4] Y Zeng,S Yin,J Liu,M Zhang,"Research of improved FP-Growth algorithm in association rules mining" ,Scientific Programming,**(2015)**,pp.1-6.

[5] Xiong Zhikang,"Protein classification based on graph kernel", Beijing University of Technology,**(2015)**.

[6] R Andersen, SO Gharan, Y Peres, L Trevisan, "Almost Optimal Local Graph Clustering Using Evolving Sets", Journal of the Acm 63.2,**(2016)**,pp.1-31.

[7] PK Manna,N Shah,S Chattopadhyay,"Extending Kernighan–Lin partitioning heuristic for application mapping onto Network-on-Chip", Journal of Systems Architecture,vol.60,no.7,**(2014)**,pp.562-578.

[8] Ye, Xiucai, and T. Sakurai,"Robust Similarity Measure for Spectral Clustering Based on Shared Neighbors", Etri Journal **(2016)**.

[9] Thomas L T, Valluri S R, Karlapalem K. MARGIN,"Maximal Frequent Subgraph Mining",Data Mining,ICDM '06. Sixth International Conference on. IEEE: **(2006)**,pp.1097 － 1101.

[10] Yan X, Han J, "CloseGraph: mining closed frequent graph patterns".Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, **(2003)**,pp.286-295.

[11] Myers, Richard, R. C. Wilson, and E. R. Hancock. "Bayesian Graph Edit Distance", IEEE Transactions on Pattern Analysis & Machine Intelligence ,vol.22,no.6,**(2000)**,pp.628-635.

[12] Liu Yongqiang."Research on graph matching algorithm based on edit distance graph", Xi'an University Of Architecture And Technology, **(2015)**.

[13] Shelokar,Prakash, A. Quirin, and Óscar Cordón. "MOSubdue: A Pareto dominance-based multiobjective Subdue algorithm for frequent subgraph mining", Knowledge & Information Systems vol.34,no.34,**(2013)**,pp.75-108.

[14] Bhuiyan, Mansurul A, and M. A. Hasan. "MIRAGE: An Iterative MapReduce based FrequentSubgraph Mining Algorithm",IEEE Transactions on Knowledge & Data Engineeringvol. 2,no.3,**(2013)**,pp.608-620.

[15] Washio,"Frequent subgraph discovery", **(2015)**.

[16] Yan X, Han J. g,"Span: graph-based substructure pattern mining", Data Mining, 2002. ICDM (2003)Proceedings. 2002 IEEE International Conference on. IEEE, **(2002)** ,pp.721.

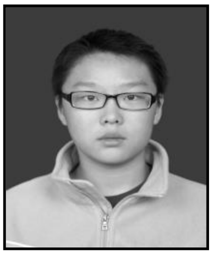
[17] Villela, Saulo Moraes, S. D. C. Leite, and R. F. Neto. "Incremental p -margin algorithm for classification with arbitrary norm." Pattern Recognition 55**(2016)**,pp.261-272.

[18] Summer,Georg."cyNeo4j-Connecting Neo4j and Cytoscape." Bioinformatics vol.31,no.23,**(2015)**,pp.3868-9.

[19] Snodin, B. E."Introducing improved structural properties and salt dependence into a coarse-grained model of DNA. " Journal of Chemical Physics 142.23(**2015**),pp.06B613_1.

[20] Shahrivari S, Jalili S."Single-pass and linear-time k-means clustering based on MapReduce". Information Systems, (**2016**), 60(C),pp.1-12.

## Authors



**Meng-ke Zhang** (1995-),male, university student, major in e-commerce and law.

**Ping-ping Wei** (1990-), female, master graduate student, research directions and integration for database mining.

**Su-zhi zhang** (1965-), male, Ph.D., professor, research direction for the Web database, distributed computing and heterogeneous system integration.

**Jia-xing xu** (1990-), male, graduate, research direction and integration for database mining.