A Novel Clustering Algorithm for Large-Scale Graph Processing

Zhaoyang Qu[1], Wei Ding[1], Nan Qu[2], Jia Yan[3], and Ling Wang[1(✉)]

[1] Department of Computer Technology, School of Information Engineering, Northeast Dianli University, Jilin, China
qzywww@nedu.edu.cn, 649993290@qq.com, smile2867ling@163.com

[2] The Jiangsu Province electric power overhauls company, Suzhou, Jiangsu, China
157729173@qq.com

[3] State Grid Jilin Electric Power Co. Ltd., Changchun, Jilin, China
407401293@qq.com

Abstract. The most important issue of big data processing is the relevance of analytical data; thought of this paper is to analyze the data as a graph optimal partitioning problem. Computing all circuit graphics firstly, calculated frequent map and redrawing of the system structure according to the results, the core problem is the time complexity of the algorithm. To solve this problem, researching DEMIX algorithm in non-strongly connected graph and study on relationship between frequent node and adjacency matrix which is strongly connected branches. Gives the corresponding examples, and analyzes the algorithm complexity. On the time complexity of the proposed method DEMIX is retrieving effect faster, more accurate search results.

Keywords: Large data · DEMIX algorithm · Graph

1 Introduction

With the popularization of Internet and Big Data, used graph theory to deal with the growing trend of data structure becomes currently research. Mining frequent sub-graphs is an important operation on graphs [1]. Most existing work assumes a database of some small graphs, but modern applications, such as smart grid, sensor networks and energy Internet [3], etc. Intellectual grid is a combination of integrated information, communicated technology, transmission and distributed infrastructure, which is quite obviousness that large graphs constitute millions of vertices and more than tens of millions of edges, as shown in Fig. 1. On the whole, we desired for establishing a well-built graphical structure or design an algorithm for retrieve and update data more efficiently.

In a big data diagram, the method of research on a large graph has become a hot issue, it's axiomatic that more difficult to analyze graph data or to optimize the structure for large data fields, due to the structure of a data complexity and size. User always search result cannot guarantee the most relevant information; the low accuracy rate

D.-S. Huang et al. (Eds.): ICIC 2016, Part III, LNAI 9773, pp. 349–358, 2016.
DOI: 10.1007/978-3-319-42297-8_33

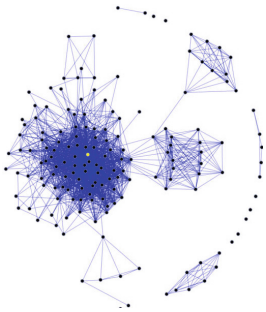

Fig. 1. Large data graph

reduces the user's experience. For large data graph, sub-graph matching is a very important skill. However, only by matching pairs figure does not solve the problem of improving efficiency of update and shortening high-quality search time pursuit.

The emphasis is aiming to make a better consequent of the problem at current circumstances in paper; it turns out working on solving the smart grid energy internet etc., particularly graph G of big-data and this sub-graph, they are possibility constituted of multi-million of points and edges, which are applied to sub-indexes that indexed possible relationship from among all peaks. Notwithstanding, in earlier work, correlated characteristic figured out problem of update searching. On the other hand, we attempted at figured out a model of a hierarchical structure that extracted valuable data, which disassemble complex relationship of big-data diagram and integrated it. Therefore, we devoted ourselves to lessening cost time of architecture of system, which efficiently researching large data set by designed efficient algorithm and structures of index.

2 Related Work

Seeking for large data diagram of smart grid, the key to building a search relevant sub-graph, which, in this paper, speedily searching and accurately matching of graph. There are some work-related studies demonstration that searched by time complexity is more efficiently than the whole graph.

American Electric Reliability Technology Association (CERTS) funded the development of DER-CAM can be micro-grid in the lowest cost for energy or CO2 emissions as low as optimization target plan single or multiple targets can be determined inside the micro-grid distributed energy excellent combination of capacity and the corresponding operational plans [4]. Hybrid Optimization Based on Genetic Algorithm Design Software (HOGA) Department of Electrical Engineering University of Zaragoza in Spain developed by a single goal or objective can be optimized [5]. Japan's Kyushu Electric Power company uses Hadoop cloud computing platform for the mass consumption of electricity user system data for rapid analysis, and the development of various types of distributed batch processing applications in the platform on the basis of improved data processing speed and efficiency [6, 7]. Thereby

improving the search graph pruning effect; MQC algorithms for real-time data to a relatively short period property which help to build the vertex set and edge set. Along with structure points and edges of the graph contained in time series order, the algorithm can visually describe events in real-time to change circumstances in the process [8]; effectively enhance the accuracy and usefulness of the real-time graph. However, in the majority of the data set, the job or task usually has a potential link, thus it is very important in the process of updating and searching when we go through the traditional method to traverse the entire data setting which consumes huge. Although the above method to search and match data is a good approach [9], our updating and searching from the perspective of an approaching relevance to the entire system is a more like a hierarchical diagram conserving system resources, which will figure out the specific relevant naphthalene extraction and build frequent node set, thus greatly improve the efficiency of the search and update.

In the current popular large graph map update process, we are more concerned about how to design a real-time updates algorithm, in other words, process the dynamic updates of data structures. In the real time case, a relatively small amount of data, a time or event nodes balanced distribution system updates the progress effectively will save the consumption of the system, which facilitates the attention of the most popular, the hottest, most valuable topics or events [10–15]. However, due to the increasing amount of data, from the perspective of the future development of the amount of data we are looking at now, perhaps, what we now call the Big Data is only "small data". Therefore we need to update large data structure diagram quickly which we could use in study and explore various methods, such as design a naphthalene structure searching algorithm [16–20]. In the above example, we point out: There is an active node, through which we pass messages very quickly, this indicates that if we go to an event updates via the node groups which are also concerned will keep that situation, Extracting these nodes are often very necessary. At the same time, and it exists in the relationship between the group closed loop, on behalf of certain events focus groups among them, so extracting the key part of these comparative studies of valuable information really matters.

This paper principally focused on studying other method of processing diagrams, discussing their limitations, studying the structure of subgraph, analyzing the correlation relationship which including relational model that building the graph. Finally, we proposed the core methods in this paper.

3 Strongly Connected Graph Structure

In this chapter, the problem is defined; well the primary problem can be solved the investigation of data graph. We present the views about how to build strongly connected structure in our algorithm, which need be satisfied to optimize searching structure and to update purposes. Strongly connected components are actual connection which studied in this paper. Particular, to avoid shortcomings in the search and update, we need to define a few attributes.

Definition 1 (Frequent Node). *Give a large graph SG, if node d have many properties, we should make sure that there is a value introducing graph system structure design, therefore, for the large graph, there is intermediate node d in the layer memory SG, i.e. d∈SG, d is greater than the threshold value of the point.*

Definition 2 (Loop and Trajectory). *Given G, a trajectory X is a sequence ((x1, x2), (x2, x3),..., (xk, x1)) such that there exists a path x1 → x2, → ,..., → xk → x1 on G, we recorded and marked down for this track. Besides, if an x1–xn path is a nonempty graph X.s = (Vi, Ei), and Vi = {x1, x2...xn} and Ei = {(x1, x2),..., (xn − 1, xn), (xn, x1)}, so X.s is a sub-graph of G and the path are not identical.*

Definition 3 (Vertex Frequency). *Given G, vertex set V(i) ∈ G,we distinguish by calculating the thresholds Θ to determine which vertices may be added to the data set SG, considering the follow-up study. In* Fig. 2*, for example, dotted line is an infrequent vertex while solid line is strong ties, so the strong ties vertexes belong to naphthalene structure graph.*

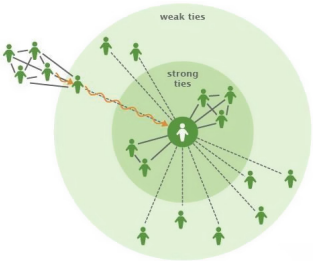

Fig. 2. Strong ties and weak ties

A formal definition should be proposed to facilitate the understanding, here are the main notation used in this paper and the rest of the column in Table 1 we used to explain implication.

According to the above definition clearly expressed, it is the main concept of this paper. The following part will describe how to create a strongly connected graph, as well as implementation of the algorithm.

Table 1. Notations

Notation	Description
G	*Large graph*
SG	*Strongly connected graph*
V(*)/E(*)	*Vertex/edge set of **
X	*Index graph*
X.s	*Meet the conditions index graph*
d(*)	*Frequency of vertex set of **
α	*Frequency of X.s*
β	*Weakly connected graph*
Θ	*Analyzing critical value*

Graphics extract strongly connected structure is the core of DEMIX algorithm. From the performance point of view, frequent node needs to be extracted, because they have a separate high-value associated with the data, which are stored separately and found a similar pattern by looking for contacts and information invisible loop when large data exists. Therefore, this paper provides a construct to calculate the first major design DEMIX graph each point must be determined according to the degree of connectivity frequently find the figure all the loops, generated between the cross find all strongly connected components in graph. If it is lower than the threshold by nodes frequency, not extracted it. If opposite, marking nodes and node records related trajectories.

4 DEMIX Clustering Algorithm

In this section, diagram cluster algorithm DEMIX will be introduced. DEMIX with trim and matching algorithm starts, dealing mainly with issues strongly connected components. DEMIX There are three main steps. First DEMIX Search, and then build a hierarchical sub-graph of the last indexing structure: a sub-query in the graph G.

DEMIX clustering algorithm is a graph theory, it belongs to the category of blind search, which aims to expand the institutional, aims to identify the results of its examination of all the nodes in diagram. The goal which we want to achieve is to handle large data graphs. Each loop represents a group of followers of the event; the results of the first part of the data processing will be the next steps for the polymerization.

4.1 Constructing Stratified Sub-graph

In order to better and faster search and update the entire graph structure, the aim is to construct a sub-graph (SG). The main task is based on the strong graph search system architecture, in order to meet the data to build a secondary threshold criteria and the adjacent node graph closely integrated. Layered system concept mentioned represents an event or joined together collections of other relationships.

In fact, it is necessary to extract the index filter layer data for large data graph nodes frequently set the threshold Θ. Algorithm settings α conditions such frequent node hash table stored in the table; meanwhile, we want to establish a good mapping nodes, and by a weak connection diagram β relationships with other node set in large data graph, we ignore the relationship between a specific point to point represents, on the contrary, the form of the entire extracted storage node list. Because of the relevance of the data extremely strong, so this represents a good structure of this paper studies the value of the smart grid, as long as we judge frequently loop nodes, higher demands on the efficiency of the system can be realized.

4.2 Algorithm

The main indicators of the study are based on the structural strength of the connector assembly, the main indicators index layer, and the first layer is supplemented by the index. Objective to construct the index structure to make it easier and faster access to the data graph, the index is mainly based on the frequent node α stored in a hash table, because the loop is formed by a plurality of nodes, each node has its own letters. Naphthalene map data set mapping structure α and β have also maintained the original relationship between β and the first layer nodes. Frequent node data set α is uniformly distributed, this expression is the key nodes, where α is a key indicator of the first layer and the contact layer data structure, when the search index layer structure cannot get the right results, the search to express the α the first layer, the results continue to match the potential to meet the needs of users. For clarity, the main steps used in this paper are presented in Table 2.

Table 2. Main steps for the DEMIX

```
1.  Input :Ω |V (G)|   |V (G)| matrix, SG;
2.        for each visit[v]  1 to n;
3.            do map[v] > = u;
4.               v   Q returned queue array by calling DEMIX (G,
    map);
5.                  G^T   TRANSPOSE-GRAPG(G);
6.               w   returned queue array by calling DEMIX (G^T,
    map);
7.        for each u ∈ SG do
8.               adj^T[u]  NIL
9.        for each u ∈ SG do
10.          for each v ∈ adj[u] do
11.              INSERT(adj^T[v], u)
12.              repeat steps;
13. return SG;
```

5 Experimental

In this section, we analyzed the performance of our method comparing SAPPER by experiment [6]. SAPPER is a very important method for sub-graph matching in graphs of large size due to its excellent timeliness.

In this part, we selected twitter real data-set experiments and detailed data are shown in Table 3 (http://snap.stanford.edu/data/egonets-Twitter.html), data set includes more than 81,000 nodes and 1,768,000 edges, and the experiments used Java programming language based on Vulcan 64G memory graphics processor.

First, since the data set involved is a Stanford University tidied large graph data, experiments aim mainly to compare through the following sections: algorithm query time and the accuracy of the data. As shown in Fig. 3, the comparison index size at the different number of index nodes, Because of advantages in the naphthalene graphical

Table 3. Dataset

Parameter	Default value
Number of peaks in G	81306
Number of edges in G	1768149
Nodes in largest WCC	81306 (1.000)
Edges in largest WCC	1768149 (1.000)
Nodes in largest SCC	68413 (0.841)
Edges in largest SCC	1685163 (0.953)
Average clustering coefficient	0.5653
Number of triangles	13082506
Diameter (longest shortest path)	7
90-percentile effective diameter	4.5

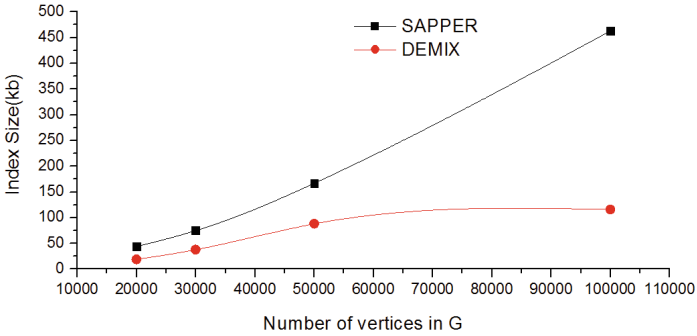

Fig. 3. Index size (Color figure online)

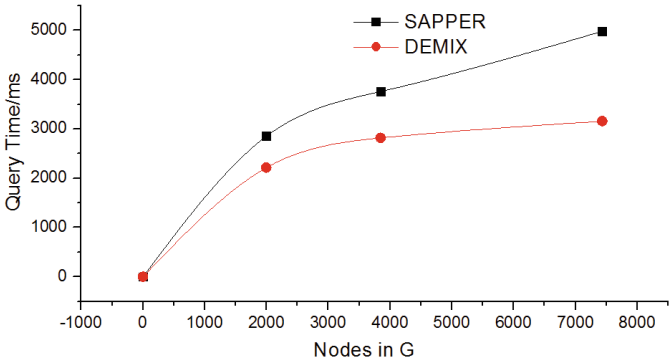

Fig. 4. Query time by different nodes (Color figure online)

structure, DEMIX algorithm has better performance than SAPPER obviously after 5000 nodes as seen from the figure; in Figs. 4 and 5, Comparing the query time by different number of node or edges, and in smaller data sets, this method showed no advantage, but when the number of data increased, our algorithm is much better than SAPPER, because many edges have been cut through DEMIX, reduced dataset infrequently expenses when searching. Figure 6 is the algorithm accuracy in different size of the data sets.

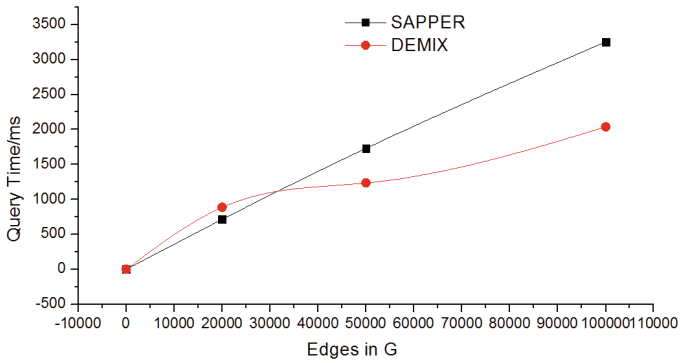

Fig. 5. Query time by different edges (Color figure online)

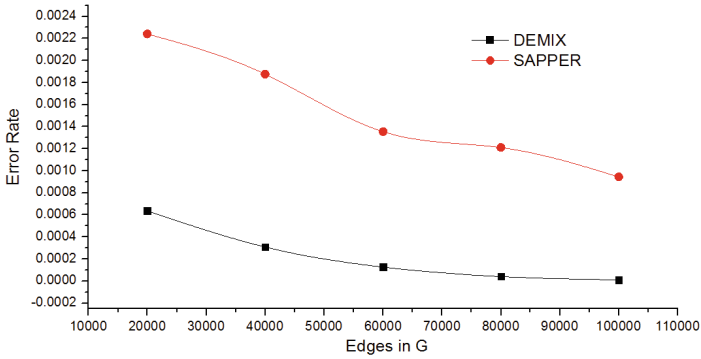

Fig. 6. Data set error rate (Color figure online)

6 Summary

In this paper, we discuss the construction of a large graph diagram in the DEMIX structure model through introducing naphthalene diagram design aiming to achieve fast search and effective algorithm updates. Experimental results show that the method of the research for graph computing that exists in the cluster event concerned and fixed interest groups such as messages delivering faster than conventional sub-graph

matching method has great research value. In the following study, we will continue to research the potential of biphenyl existing in a smart grid, and the construction of multilevel, high sensitivity graph structure.

Acknowledgements. This work is supported by the development of National Natural Science Foundation Project (No. 51277023), by the Jilin Province plans to emphasis transformation projects (No. 20140307008GX), and by the Education Department Foundation of Jilin Province (No. 201698).

References

1. Elseidy, M., Abdelhamid, E., Skiadopoulos, S.: GraMi: frequent subgraph and pattern mining in a single large graph. Proc. VLDB Endow. **7**(7), 517–528 (2014)
2. Zhou, Y., Cheng, H., Yu, J.X.: Graph clustering based on structural/attribute similarities. Proc. VLDB Endow. **2**(1), 718–729 (2010)
3. Zhao, F., Tung, A.K.H.: Large scale cohesive subgraphs discovery for social network visual analysis. Proc. VLDB Endow. **6**(2), 85–96 (2012)
4. Martinez, C.: Intelligent real-time tools and visualizations for wide-area electrical grid reliability management, pp. 1–4 (2008)
5. Zhang, Y., Zhang, H., Lu, C.: Study on parameter optimization design of drum brake based on hybrid cellular multiobjective genetic algorithm. Math. Probl. Eng. **2012**(1), 1–18 (2012)
6. Miyake, Y., Tanaka, K., Okubo, H.: Seaweed consumption and prevalence of depressive symptoms during pregnancy in Japan: baseline data from the Kyushu Okinawa maternal and child health study. BMC Pregnancy Childbirth **14**(5), 572–578 (2014)
7. Williams, M., Wallis, S., Komatsu, T.: Dragons, brimstone and the geology of a volcanic arc on the island of the last Samurai, Kyushu, Japan. Geol. Today **32**(1), 21–26 (2016)
8. Li, Y., Liu, Z., Zhu, H.: Enterprise search in the big data era: recent developments and open challenges. Proc. VLDB Endow. **7**(13), 1717–1718 (2014)
9. Agarwal, M.K., Ramamritham, K., Bhide, M.: Real time discovery of dense clusters in highly dynamic graphs: identifying real world events in highly dynamic environments. Proc. VLDB Endow. **5**(10), 980–991 (2012)
10. Gupta, P., Satuluri, V., Grewal, A., et al.: Real-time twitter recommendation: online motif detection in large dynamic graphs. Proc. VLDB Endow. **7**(13), 1379–1380 (2014)
11. Pavan, A., Tangwongsan, K., Tirthapura, S., et al.: Counting and sampling triangles from a graph stream. Proc. VLDB Endow. **6**(14), 1870–1881 (2013)
12. Budak, C., Georgiou, T., Agrawal, D., et al.: Geoscope: online detection of geo-correlated information trends in social networks. Proc. VLDB Endow. **7**(4), 229–240 (2013)
13. Wu, X., Zhu, X., Wu, G.Q., et al.: Data mining with big data. IEEE Trans. Knowl. Data Eng. **26**(1), 97–107 (2014)
14. Cohen, J., Dolan, B., Dunlap, M., et al.: MAD skills: new analysis practices for big data. Proceedings VLDB Endow. **2**(2), 1481–1492 (2009)
15. Agarwal, M.K., Ramamritham, K., Bhide, M.: Real time discovery of dense clusters in highly dynamic graphs: identifying real world events in highly dynamic environments. Proc. VLDB Endow. **5**(10), 980–991 (2012)
16. Kim, Y., Moon, J., Lee, H.-J., Bae, C.-S.: Knowledge Digest Engine for Personal Bigdata Analysis. In: Park, J.H(., Jin, Q., Yeo, MS.-s., Hu, B. (eds.) Human Centric Technology and Service in Smart Space. LNEE, vol. 182, pp. 261–267. Springer, Heidelberg (2012)

17. Hoff, P.D., Raftery, A.E., Handcock, M.S.: Latent space approaches to social network analysis. J. Am. Stat. Assoc. **97**(97), 1090–1098 (2002)
18. Valente, T.W.: Social network thresholds in the diffusion of innovations. Soc. Netw. **18**(1), 69–89 (1996)
19. Wang, W., Yang, J.: Mining sequential patterns from large data sets. Adv. Database Syst. **28** (7), 3–14 (2013)
20. Bullmore, E., Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci. **10**(3), 186–198 (2009)