Development of Efficient Network Analysis System for Revealing Potential Trade Risk Factors in *e*-Customs

Dongmin Seo and Min-Ho Lee

***Abstract*—Recently, various network analysis methods has been utilized to reveal undisclosed knowledge in a variety of fields. In particular, these methods are used to reveal potential trade risk factors in *e*-Customs. However, existing methods do not provide a fast response time to user queries, mainly due to the large size of the data and the complexity of relationships between the data in *e*-Custom. In this paper, we propose an efficient network analysis system for revealing potential trade risk factors in *e*-Custom. The system proposes an efficient subgraph matching method and visualization tool to find the relationships between the data in a network. It quickly finds complicated relationships and dramatically reduces the number of unnecessary searches. Also, to verify the superiority of our method, we compare our method with existing method in various experiments.**

***Index Terms*—Network analysis, subgraph matching, e-customs, big-data.**

I. Introduction

Big data refers to informationalization technology for extracting valuable information through the use and analysis of large-scale data and, based on that data, deriving plans for response or predicting changes. As an example, the rise of petabyte scale data warehouses, social networks, real-time sensor data, and diverse and new data sources has led to the ability to address many problems. In addition, with the rise of research environments centering on data, interdisciplinary cooperation has increased [1].

Big data is composed of large networks. For example, the relations between friends in Facebook and the relations between compounds, genes and proteins in a brain are networks. A network is a data structure that represents the relations between data and is consists of nodes and edges. If a network structure is analyzed, it can discover characteristics and patterns between data in a network. For example, a product recommendation that recommends the most similar item purchased by similar people, an influencer identification that finds out people that are central in the given network, community detection that identifies group of people that are close to each other and graph pattern matching that finds out all the sets of entities that match to the given pattern are based on a network analysis. Recently, a graph DBMS is growing rapidly in a DBMS market because most big data analyses are based on network analyses, as shown in Fig. 1. The graph DBMS is growing at an average annual rate of 40.8% [1].

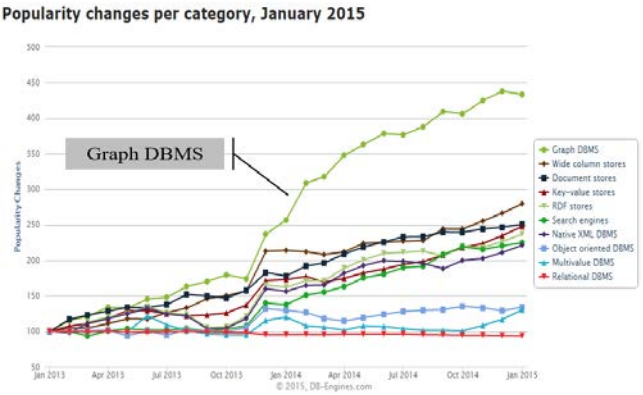


Fig. 1. Graph DBMS market trend.

Along with the development of globalization and information technology, the trade system has changed from international trade to free trade. The development of free trade has caused an increase in the trade volume and an increase in the number of risk factors. Furthermore, due to the proliferation of trade and travelers, the tasks of customs services are steadily increasing. Consequently, an efficient information analysis system is required to detect and prevent illegal acts such as terrorism, smuggling, and trade in hazardous substances. To this end, the Korea Customs Service (KCS) developed *e*-Customs by using advanced IT technology in 2008 [2]-[4].

e-Customs stores information about travelers, posts, trade facilitation and community protection while maintaining an information plaza with a variety of real-time information in relational databases. It also supports an information analysis service. However, *e*-Customs can only answer queries about intuitive knowledge such as “Who boarded an international flight yesterday?” and “What was the direct transactional information between two companies last month?” because the information analysis service of *e*-Customs does not reveal the complex relationships between data in relational databases [4], [5]. Therefore, customs inspectors require an advanced information analysis service to track all relationships between potential trade risk factors efficiently.

In this paper, we propose an efficient network analysis system for revealing potential trade risk factors in *e*-Customs. Our system provides an efficient subgraph matching method and visualization tool. The proposed subgraph matching method quickly finds various and complicated relations because it dramatically reduces the number of searches for unnecessary relations.

The remaining part of this paper is organized as follows. In Section II, we explain related works. In Section III, we

Manuscript received February 7, 2018; revised March 23, 2018. This work was supported by the K-18-L03-C02 and K-18-L11-C03 funded by Korea Institute of Science and Technology Information.

Dongmin Seo and Min-Ho Lee are with the Korea Institute of Science and Technology Information, Daejeon, Republic of Korea (e-mail: dmseo@ kisti.re.kr, cokeman@ kisti.re.kr).

doi: 10.18178/ijfcc.2018.7.1.511

describe the subgraph matching method and visualization tool of our system. In Section IV, we show our experimental results briefly. Finally, we offer a conclusion for this paper in Section V.

II. Related Works

A. Subgraph Matching

A subgraph matching is the process of finding a correspondence between the nodes and the edges of two networks that satisfies some constraints ensuring that similar substructures in one network are mapped to similar substructures in the other [6]. The subgraph matching is widely used in cheminformatics, bioinformatics, image processing and computer vision and social networks analyses. However, the existing subgraph matching methods are unsuited for large networks. For example, VF2 [7] is the most representative subgraph matching method and is the fastest subgraph matching method on small networks. However, VF2 is unsuited for large networks and has an overlap problem on subgraph matching results, as shown in Fig. 2. Also, the overlap problem of VF2 occurs unnecessary graph traversals [8]. Therefore, we proposed an efficient subgraph matching on large networks in this paper.

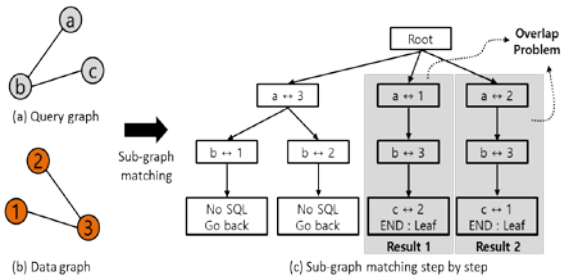
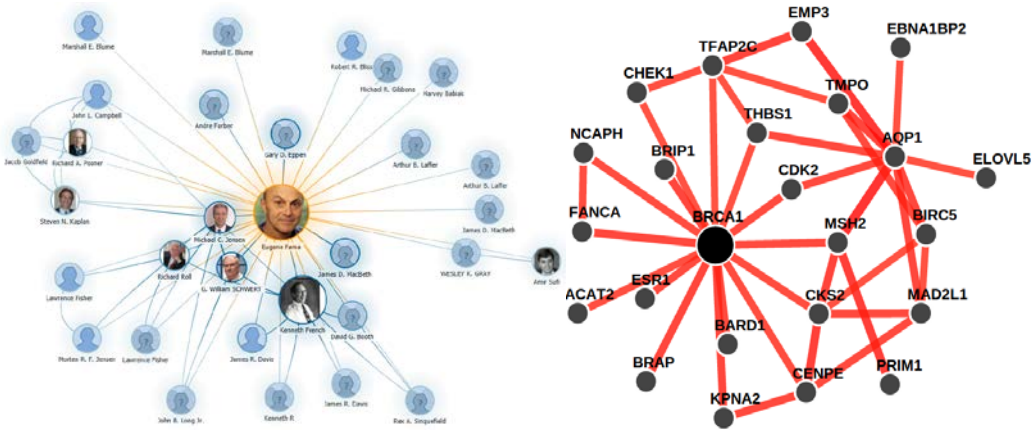


Fig. 2. The example for VF2.

B. Information Retrieval Service Based on Relationships between Data

Fig. 3 shows systems that provide specialized information analysis based on the search for relationships between data. Fig. 3 (a) shows Microsoft's co-author graph service [9], which has been in service since 2008, and provides all the author information related to two authors based on the author's information of the papers and patents. Fig. 3(b) shows the Gene network service [10], which has been in service since 2010, and provides a network of other proteins related to the protein inputted by a user. Recently, information retrieval services based on relationships between data have been utilized in various fields. Also, the services are saving time and effort in analyzing information.



(a) Co-author graph service (b) Gene network service

Fig. 3. Typical search services to find relationships between data.

III. Proposed Network Analysis System for *e*-Custom

A. Proposed Subgraph Matching

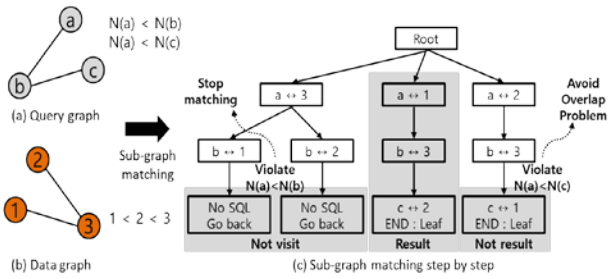


Fig. 4. The example for proposed method.

Fig. 4 shows our subgraph matching method. Our method assigns the priorities of the nodes in a data graph and query graph. In a data graph, the node with the smallest number of degrees has the highest priority. The priorities of nodes with the same number of degrees are random. In Fig. 4, node 1 has the highest priority and node 3 the lowest priority. In a query graph, the priorities of nodes are based on the order of node traversal. In Fig. 4, nodes are traversed in order of node a, node b and node c. Therefore, the priorities of the query graph are node a < node b and node a < node c. Node a < node b means that the priority of node a is higher than that of node b. For lack of space, we won't deal with the detail description here. Fig. 4 shows the process of our subgraph matching. The subgraph matching based on the priorities of nodes avoids an overlap problem on subgraph matching results. For example, after node a and node 2 are matched, node b and node 3 are matched because the priority of node a is higher than that of node b and the priority of node 2 is higher than that of node 3. However, node c and node 1 aren't matched because the priority of node a is higher than that of node c but the priority of node 2 is lower than that of node 1. In conclusion, (a↔2, b ↔3, c↔1) isn't the result of the subgraph matching. Also,

our subgraph matching is quicker than VF2 because the subgraph matching based on the priorities of nodes avoids unnecessary graph traversals [8].

B. Proposed Visualization Tool

The data used in *e*-Customs are security data. So, Fig. 5 shows the dummy data that are used for revealing potential trade risk factors in *e*-Customs. Referentially, the *e*-Customs uses more data, but Fig. 5 only shows key data for revealing potential trade risk factors. For example, the sample-02 record means that Gil-dong imported cell phones from the company B. The origin of the cell phones is VN(Vietnam), and it was imported from the RS(Russian Federation) to the Incheon port by the B_Express. The cell phones were notified of import through A_T.A(A tax accountant). Also, the cell phones were inspected for origin violations, but there were no problems. On the other hand, the sample-07 violated a origin and the sample-09 violated a price report. The 'null' of an inspect item code and inspect result columns means the goods are imported without being inspected. Finally, the key value is an import declaration code.

Column Name	Sample-01	Sample-02	Sample-03	Sample-04	Sample-05	Sample-06	Sample-07	Sample-08	Sample-09	Sample-10
Report Date	20171207	20171207	20171207	20171207	20171106	20171106	20171106	20171005	20171005	20171005
Import Declaration Code	0001|01	0001|02	0001|03	0001|04	0002|01	0002|02	0002|03	0003|01	0003|02	0003|03
Taxpayer Code	Gil-dong	Gil-dong	Dae-dong	Beom-soo	Gil-Dong	Gil-dong	Dae-dong	Beom-soo	Dae-dong	Cheol-soo
Item Code	Qquid	Cell-phone	Qquid	Bag	Crab	Cell-phone	Laptop	Clock	Bag	Bag
Overseas Account Code	A_Comp	B_Comp	C_Comp	D_Comp	B_Comp	B_Comp	C_Comp	E_Comp	F_Comp	E_Comp
Freight Carrier Code	A_Express	B_Express	C_Express	D_Express	B_Express	B_Express	C_Express	F_Express	F_Express	E_Express
Extracted Country Code	BN	BN	BN	CN	BN	BN	BN	CN	CN	CN
Country Code of Origin	BN	RS	BN	EG	RS	RS	BN	SG	HK	HK
Bonded Area Code	Incheon_Port	Incheon_Port	Incheon_Port	Incheon_Port	Incheon_Port	Incheon_Port	Busan_Port	Busan_Port	Incheon_Port	Busan_Port
Tax Accountant Code	A_T.A	A_T.A	B_T.A	C_T.A	A_T.A	C_T.A	B_T.A	D_T.A	E_T.A	D_T.A
Inspect Item Code	null	Origin	null	null	Origin	null	Origin	null	Price	null
Inspect Result	null	N	null	null	N	null	Y	null	Y	null

Fig. 5. The dummy data used for revealing potential trade risk factors in *e*-Customs.

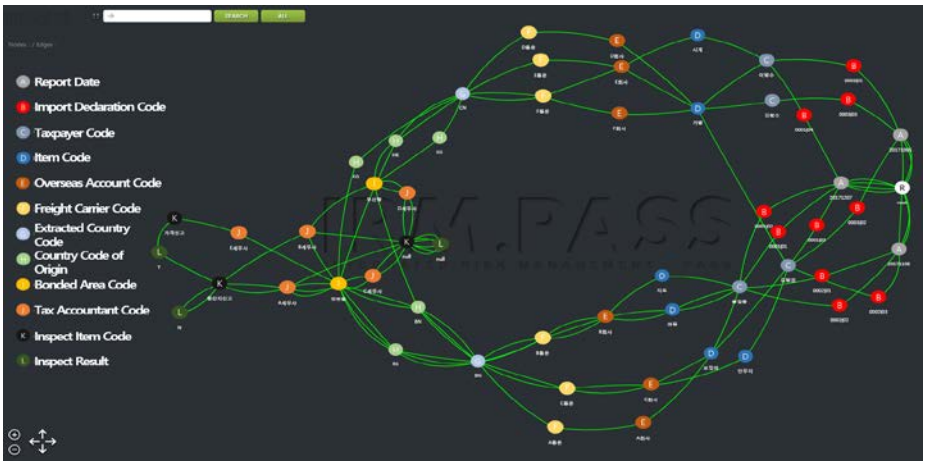
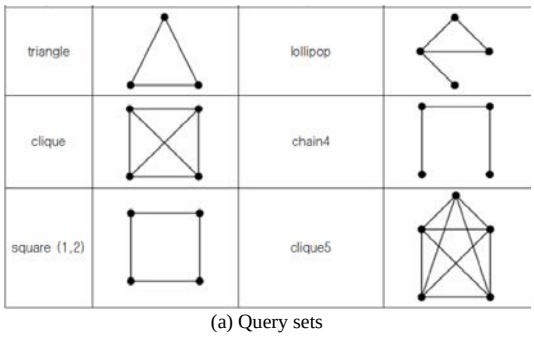


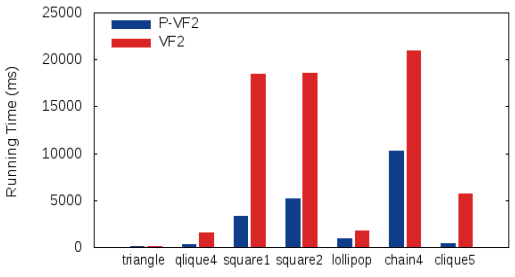
Fig. 6. Our visualization tool.

Fig. 6 shows our visualization tool that transforms and visualizes a relational database into a network. In conclusion, Fig. 6 shows a network for Fig. 5. Our visualization tool supports the various network analysis and visualization functions. The proposed visualization tool is developed by state-of-art web technology such as HTML5 and Ajax for visualizing a large network efficiently. Also, the visualization tool supports zoom in/out, gravitational constant, and node/edge hidden on a network. Then the user can distinguish complex network relationships easily.

Fig. 7. The result of a subgraph matching by our method.



(a) Query sets



(b) Running time

Fig. 8. The evaluation result on MINT datasets.

Customs inspectors look for patterns that can reveal potential trade risk factors based on existing detected trade risk factors. For example, customs inspectors guess that C_Comp, C_Express and B_T.A in the sample-07 are potential risk factors. So, the customs inspectors look for all trades that include C_Comp, C_Express and B_T.A and was not inspected.

Fig. 7 shows the result of the subgraph matching that includes (Overseas Account Code: C_Comp), (Freight Carrier Code: C_Express) and (Tax Accountant Code: B_T.A). In conclusion, Fig. 7 shows a network for the sample-03. Our subgraph matching method quickly performs complex subgraph matching queries.

IV. Performance Evaluation

In order to verify the effectiveness of the proposed subgraph matching method, we compared our method with VF2. The evaluation has been performed on MINT dataset with 3,872 nodes and 56,937 edges. Our method is more than 130% faster than VF2 because our method avoids many unnecessary graph traversals.

V. Conclusion and Future Work

The various relationship analysis methods based on a network analysis have been utilized to reveal undisclosed knowledge in a variety of fields. However, existing methods don't provide fast response time to user queries because the data and relationships between the data very large and complex. So, we proposed a new relationship tracking method based on a subgraph matching to analyze the relationships between potential trade risk factors. The proposed method reduced the number of searches for unnecessary relationships when tracking the relationships in a network. In the future, we plan to expand our service to track relationships based on the concept of advanced international information sharing.

Acknowledgment

This work was supported by the K-18-L03-C02 and K-18-L11-C03 funded by Korea Institute of Science and Technology Information.

References

[1] D. M. Seo, S. J. Yu, and M. H. Lee, "Efficient analysis and visualization system for large networks," in *Proc. International Conference on Convergence Content*, vol. 15, no. 2, pp. 299-300, 2017.
[2] Y. C. Kim, G. Y. Ru, and S. H. Shin, "An empirical study on the modelling risk management for cumstoms administration," *J. Korea Research Society for Customs*, vol. 8, no. 4, pp. 113-132, 2007.
[3] T. I. Kim and S. Y. Kwak, "A study on electronic clearance system(UNI-PASS) in Korea," *J. Korea Research Society for Customs*, vol. 9, no. 4, pp. 69-87, 2008.
[4] P. Kim, D. M. Seo, H. M. Jung, K. S. Kim, and I. C. Yun, "The relation tracking service of customs service based on semantic web," in *Proc. International Conference on e-Learning, e-Bus, EIS, and e-Gov*, pp. 554-555, 2011.
[5] B. S. Lee, "Dynamic perspective on the advancement of Korea'S electronic customs clearance system," *J. The International Commerce & Law Review*, vol. 44, no. 12, pp. 213-238, 2009.
[6] D. Conte, P. Foggia, C. Sansone, and M. Vento, "Thirty years of graph matching in pattern recognition," *J. Pattern Recognition and Artificial Intelligence*, nol. 18, no. 3, pp. 265-298, 2004.
[7] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, "An improved algorithm for matching large graphs," in *Proc. 3rd IAPR-TC15*, 2001, pp. 149-159.
[8] D. M. Seo, H. M. Park, S. J. Yu, M. H. Lee, and U. Kang, "Efficient subgraph matching on large networks," in *Proc. International Conference on Convergence Content*, vol. 14, no. 2, 2016, pp. 275-276.
[9] Microsoft Research. (2008). Microsoft academic search. [Online]. Available: http://academic.research.microsoft.com/
[10] Gene Ontology Consortium. (2010). The gene ontology. [Online]. Available: http://www.geneontology.org/

Dongmin Seo received his B.S., M.S. and Ph.D. in information and communication engineering from Chungbuk National University, Korea in 2002, 2004 and 2008, respectively. He is now the senior researcher in the Convergence Technology Research Division, Korea Institute of Science and Technology Information (KISTI), Korea. His main research interests include network analysis, semantic-web, moving objects database (MOD) system, wireless sensor networks (WSN), XML database system, e-Spine and biomedical convergence technology.

Min-Ho Lee received his B.S., M.S. and Ph.D. in computer engineering from Chungnam National University, Korea in 1998, 2000 and 2014, respectively. He is now the senior researcher in the Convergence Technology Research Division, Korea Institute of Science and Technology Information (KISTI), Korea. His main research interests include brain network analysis, medical big data analysis, and explainable artificial intelligence.