Interdependence between Structural Engineering and Construction Management
Edited by Ozevin, D., Ataei, H., Modares, M., Gurgun, A., Yazdani, S., and Singh, A.

ISBN: 978-0-9960437-6-2

IDENTIFICATION OF CRITICAL INFRASTRUCTURE USING DATA MINING TECHNIQUES

SU-LING FAN[1], CHANG-SAAR CHAI[2,], and KUMAR VIKRAM[1]

[1]*Dept of Civil Engineering, Tamkang University, New Taipei City, Taiwan*
[2]*School of the Built Environment, University of Reading Malaysia, Johor, Malaysia*

Critical Infrastructure (CI) is a term used to describe important national assets for producing or distributing a continuous flow of essential goods or services. They are marked by immense complexity, characterized predominantly by strong intra and interdependencies as well as hierarchies. These interconnections take many forms, including flows of information, shared security, physical flows of commodities, and others. Previous research has illustrated the relationship between the physical impacts of natural disasters and the social and economic factors on CI. Some research emphasized more the role of CI interdependencies and their importance and influence over the functioning of industries while others have looked the impacts due to disruption of CI after disasters. Nowadays comprehensive identification of all interdependency relationships of CI remains a challenge. As the complexity and interconnectedness of a country's CI evolve, threats and vulnerabilities increase. Thus, investigating how a set of CI interacts and identification of criticality of CI becomes an important topic. This research has made utilization of data mining techniques and proposes a method to identify the criticality of Critical Infrastructure so that to develop better disaster protection and prevention management.

Keywords: Critical infrastructure interdependency, General sequence pattern mining, Python.

1 INTRODUCTION

Critical Infrastructure (CI) are referring to the systems and assets, whether physical or virtual, so important to a particular country that the incapability or destruction of such systems and assets would have a debilitating impact on the security, national economy and public health and safety. These assets include, but are not limited to, facilities for transportation, communication systems, electric power systems, gas and oil storage facilities and pipelines, and government services (Chou and Tseng 2010).

Critical Infrastructure Interdependency (CII) defines the relationship between two elements of critical infrastructure as a bidirectional relationship in which the output of one is essential as the input of another (Rinaldi *et al*. 2001). Furthermore, these CIIs often create complex interdependency relationships that across system boundaries (Haimes 2005). As the complexity and interconnectedness of a country's critical infrastructure evolve, threats and vulnerabilities increase. Recently, it comes across many cases in which huge losses occur due to domino effects from some system failure. A typical example from a technical glitch of Taiwan Tatan Power Plant in 2017 which caused malfunctioning of six generators affecting the supply of 4 million

KWs of electricity. This incident brought a big impact to the local authority in transportation, manufacturing factories, hospitals, retails and many others (Chung-Han and Power 2017). Due to the great economy loses resulted from this incident, Taiwan National Security Office is prominently aware of the importance of CII in the context of homeland security. Therefore, a series of CII research and investigations are carried out to enhance the current performance (Chou and Tseng 2010).

This study is aimed to investigate the causal relationship of CII using data mining techniques. In order to achieve the aim, a case study using hospital system data were analyzed through General Sequencing Pattern and Python.

2 CRITICAL INFRASTRUCTURE INTERDEPENDENCY (CII)

There are different approaches demonstrating a CII relationship. For instance, Mendonça and Wallace (2006) employed a structure to represent the number of intersection between two critical infrastructure systems. In contrast, Mcdaniels *et al.* (2007) collected media reports and official ex-post-facto assessments to characterize the consequences of infrastructure failure interdependencies, in terms of impact and extend indices. They defined a CII relationship as a combination of its impact and extent indices. These indices can be further extended to impact score and extend score. What has heretofore been missing is an investigation of information contained in failure records. Such time-oriented failure for the Critical Infrastructure may be very difficult; using only the sequential information in the failure records may be a feasible alternative for analyzing CII relationships.

Other researchers have been concentrated efforts on the analysis of CII. Haimes and Jiang (2001) proposed the Leontief-based input-output model to investigate interconnectedness among various sectors of critical infrastructure. Although inability of one sector, called inoperability, can be precisely calculated, this model requires an input of a matrix structure denoting the relationship between each sector of critical infrastructure, and methods regarding how to generate the matrix have not been elaborated well in literature. Shih *et al.* (2009) used data warehousing and visualization techniques to explore CII for U.S. power generation. The study proposed a process to build a data warehouse, integrating data mining and geographic information system techniques so as to identify vulnerabilities in coal delivery to power plants. The results help to locate the need for redundancies in the coal supply chain.

3 BIG DATA

The Big Data concept is now receiving remarkable attention for tacking complex engineering problems. Among different fields, Big Data analytics is notably impacting the Civil Engineering domain. Currently, the operation and maintenance of Civil Engineering systems are undergoing a noticeable transformation as a result of huge amount of information provided by the emerging testing and monitoring systems. The key role of Big Data in this transformation is well understood. Despite the significance of the Big Data technologies to process large-scale data, current Civil Engineering information systems are still lacking in the successful implementation of them (Alavi and Gandomi 2017). The world is currently inundated with data, with fast advancing technology leading to its steady increase. The construction industry is not an exception to the pervasive digital revolution. The construction industry is responsible for undertaking some of the biggest and most expensive projects on Earth. Huge amounts of resources and work go into major construction projects and of course, this means that huge volumes of data are generated. The industry is dealing with significant data arising from diverse disciplines throughout the life cycle of a facility. Noticeably, this data in any form and shape has

intrinsic value to the performance of the industry. With the advent of embedded devices and sensors, facilities have even started to generate massive data during the operations and maintenance stage, eventually leading to more rich sources of Big Construction Data. This vast accumulation of Construction data has pushed the construction industry to enter the Big Data era. (Bilal *et al.* 2016).

3.1 Data Mining

Data mining is the process of finding useful information from a large amount of data. The interesting patterns can be mined with the help of the several data mining techniques by (Suguna and Nandhini 2015). Data mining is defined as the nontrivial extraction of implicit, previously un-known and potentially useful information from data (Kreng and Yang 2013). Data Mining is concerned with the automatic or semiautomatic exploration and analysis, of large volumes of data, to discover meaningful patterns or rules. Data Mining has a broader scope than other traditional data analysis fields (such as statistics) since it tends to answer non-trivial questions (Wang and Leite 2013). For patterns discovery and extraction, Data Mining is primarily based on the technique(s) from statistics, machine learning, and pattern recognition (Linoff and Berry 2011). Several models are created and tested to assess the suitability of a particular technique(s) for solving the given business problem. Models with the highest accuracy and tolerance are chosen and applied to the actual data for generating predictive results (including predictions, rules, probability, and predictive confidence) (Bilal *et al.* 2016).

The development of Big Data analysis is eventually improving. However, General Sequencing Pattern (GSP) remained as one of the best technique concerning with finding statistically relevant patterns between data examples where the values are delivered in sequence.

3.2 General Sequences Pattern (GSP)

The sequential pattern is a sequence of item sets that frequently occurred in a specific order, the items in the same item sets are supposed to have the same transaction- time value or within a time gap. Usually, all the transactions are viewed as a sequence, usually called X-sequence (X can be renamed accordingly), where each transaction is represented as an item sets in that sequence, all the transactions are a list in a certain order with regard to the transaction-time. Sequences are important data, which occur frequently in many fields such as medical, business, financial, customer behavior, educations, security, and other applications.

Several algorithms have been devised for discovering such sequential patterns. The GSP algorithm was proposed by (Agrawal and Srikant 1994). The study aimed at mining sequential patterns over a large database of customer transactions. The GSP algorithm was subsequently applied in discovering various rules underlying a given set of sequences to predict a plausible sequence continuation.

Frequently occurring patterns ordered by time are found by sequential pattern mining. Sequential pattern mining has wide application since the data has a time component attached with them. For example, the medical domain can determine a correct diagnosis from the sequence of symptoms experienced; over customer data to help target repeat customers and with web-log data to better structure a company's website for easy accessibility of most popular links. There are several known methods for discovering general sequential patterns at present. Still, in a specific domain of web-log analysis more methods exist.

4 RESEARCH METHODOLOGY

Inoperability Input-Output Model (IIM) algorithm is adopted in this study. The algorithm has been designed to determine the criticality of the system and interdependencies within the system. It uses the sequences dataset generated to compute the matrix. With the power of enormous libraries contained in Python, it produces a quantitative output in the form of a matrix generated from the relationships obtained from the sequences. Through the CII matrix, the systems, which have the higher risks of failure can be identified and precaution measure could be proposed. This matrix also helps in determining how other systems behave if one of the systems fails, subsequently explains the impact of one system to another.

The data is collected from hospital Taipei Taiwan. Data format is in MySQL (.mdf) format of size 7.09 GB, contains 50,117,166 numbers with 30 days of data. Bad data contains 137,236, which falls in 536 system. The data information contains system details, IP address, date and time and system status (good or bad).

The collected data is fall in medium scale size, therefore Python is utilized in this study. It is a fast data mining approach, which is capable to create new outcome through the analyzed results. Python has emerged as a good option in data processing, and there is often a trade-off between scale and sophistication. Besides, it has excellent amounts of toolkits and features that gave advantage on data community.

5 RESULTS AND DISCUSSIONS

The collected data is first converted to ".csv" format then imported to the MySQL server in visual studio code. The excessive rows and columns will be deleted in this process to ensure that the data is clean and fit to analyze. Next, the data is filtered through "good" and "bad" status. The bad status data will be identified to further analyzed. In order to enhance the data analyzing the process, the distinct system names are renamed to "S01", "S02" accordingly as shown in Figure 2. Finally, the data will be sorted according to time frames (1 day, 5 days and 30 days time frame) and data with same time frames will be regrouped.

After the data screening process, min_count argument method is used to filter the low probability of occurrence dataset. This is to minimize results error due to the low occurrence rate. Then, the IIM matrix method is utilized by giving different time difference of failures and the added sequences were updated to the array containing all the sequences. The repeating systems within the same sequence will be eliminated. The formulated matrix is as in Eq. (1). The value of i^{th} row and j^{th} column of the matrix was calculated as:

$$A_{ij} = \frac{count(sys(i) \cap sys(j))}{count(sys(i))} \tag{1}$$

Where A_{ij} denotes the element corresponding to the i^{th} row and j^{th} column and count is used to calculate how many times the specific system has occurred in a sequence.

The result of day 3 is shown in Figure 3. The results are in the form of a CII matrix that represents the interdependencies between systems. From this matrix, the most critical system can be found out. The element (i, j) of the CII matrix indicates the dependent level of system i on j and impact of the system i. For example, in Figure 3 the element (2,4) indicates that the system "S04" will get 15% damaged if system "S02" is damaged. In a similar way, the entire matrix has been formulated. From the matrix, it can be observed that system "S01", "S02" and "S03" are highly interdependent on each other (100% dependent) which indicates that if any of the system S01, S02 or S03 damaged, the other two systems will be definitely affected. A similar pattern is

found on systems 9 to 13 (i.e. “S09”, “S10”, “S11”, “S12”, “S13”) which are 100 percent interdependent on each other.

	System_name	ID	Date_Time	Null	IP_address	status
0	/FACILITY/E-CH-1-SV.PV	-9999	29-11-2017 09:06	NaN	192.168.0.11	OPC_QUALITY_BAD
1	/FACILITY/E-CH-1-RA.PV	-9999	29-11-2017 09:07	NaN	192.168.0.11	OPC_QUALITY_BAD
2	/FACILITY/E-CH-1-RV.PV	-9999	29-11-2017 09:07	NaN	192.168.0.11	OPC_QUALITY_BAD
3	/FACILITY/E-CH-1-SA.PV	-9999	29-11-2017 09:07	NaN	192.168.0.11	OPC_QUALITY_BAD
4	/FACILITY/E-CH-1-SA.PV	-9999	29-11-2017 09:07	NaN	192.168.0.11	OPC_QUALITY_BAD
5	/FACILITY/E-CH-1-RA.PV	-9999	29-11-2017 09:07	NaN	192.168.0.11	OPC_QUALITY_BAD

Figure 1. Data screening process for bad status data.

	System_name	time
0	S02	32760
1	S01	32820
2	S02	32820
3	S03	32820
4	S05	32820
5	S04	32820
6	S05	32820
7	S04	32820
8	S03	32820

Figure 2. System renaming.

	SYS	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	S13	S14
0	S01	1.00	1.00	1.00	0.15	0.15	0.10	0.10	0.10	0.09	0.0	0.0	0.0	0.0	0.0
1	S02	1.00	1.00	1.00	0.15	0.15	0.10	0.10	0.10	0.09	0.0	0.0	0.0	0.0	0.0
2	S03	1.00	1.00	1.00	0.15	0.15	0.10	0.10	0.10	0.09	0.0	0.0	0.0	0.0	0.0
3	S04	0.54	0.54	0.54	1.00	1.00	0.11	0.11	0.11	0.12	0.0	0.0	0.0	0.0	0.0
4	S05	0.54	0.54	0.54	1.00	1.00	0.11	0.11	0.11	0.12	0.0	0.0	0.0	0.0	0.0
5	S06	0.57	0.57	0.57	0.16	0.16	1.00	1.00	0.99	0.00	0.0	0.0	0.0	0.0	0.0
6	S07	0.57	0.57	0.57	0.16	0.16	1.00	1.00	0.99	0.00	0.0	0.0	0.0	0.0	0.0
7	S08	0.56	0.56	0.56	0.16	0.16	1.00	1.00	1.00	0.11	0.0	0.0	0.0	0.0	0.0
8	S09	0.44	0.44	0.44	0.15	0.15	0.11	0.11	0.10	1.00	0.0	0.0	0.0	0.0	0.0
9	S10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.0	1.0	1.0	1.0	1.0
10	S11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.0	1.0	1.0	1.0	1.0
11	S12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.0	1.0	1.0	1.0	1.0
12	S13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.0	1.0	1.0	1.0	1.0
13	S14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.0	1.0	1.0	1.0	1.0

Figure 3. Processed Data

6 CONCLUSION

The results that are obtained from the CII matrix are helpful in determining critical systems in the hospital. The matrix is able to predict the most likelihood of next system failure if one of the critical infrastructures is malfunctioning. This definitely helps to prevent domino effects of system failure in a particular facility. The function of this matric is not limited to compute the system in one particular facility, meanwhile, it can be extended to compute the national critical infrastructures to prevent Taiwan Tantan power plant incident to be reoccurred.

References

Alavi, A. H., and Gandomi, A. H. Big Data in Civil Engineering, *Automation in Construction*, *79*, 1–2, 2017.

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., and Pasha, M., Big Data in the Construction Industry: A Review of Present Status, Opportunities, and Future Trends, *Advanced Engineering Informatics*, *30*(3), 500–521, 2016.

Chou, C.-C., and Tseng, S.-M., Collection and Analysis of Critical Infrastructure Interdependency Relationships, *Journal of Computing in Civil Engineering*, *24*(6), 539–547, 2010.

Chung-Han, B. L. Y., and Power, K., Outage across Taiwan Due to Malfunction at Power Plant, *Focus Taiwan News Channel*, 2017.

Haimes, Y. Y., and Jiang, P., Leontief-Based Model of Risk in Complex Interconnected Infrastructures, *Journal of Infrastructural Systems*, *7*(1), 1–12, 2001.

Haimes, Y. Y., Infrastructure Interdependencies and Homeland Security, *Journal of Infrastructure Systems* 65–66, 2005.

Kreng, V. B., and Yang, S.-W., Data Mining of Hospital Characteristics in the Online Publication of Medical Quality Information, *Health*, 05(05), 931–937, 2013.

Linoff, G. S., and Berry, M. J., *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, John Wiley and Sons, 2011.

Mcdaniels, T., Chang, S., Peterson, K., Mikawoz, J., Reed, D., Empirical Framework for Characterizing Infrastructure Failure Interdependencies, *Journal of Infrastructure Systems*, 13(3), 175–184, 2007.

Mendonça, D., and Wallace, W. A., *Impacts of the 2001 World Trade Center Attack on New York City Critical Infrastructures*, 260–270, 2006.

Rinaldi, S., Peerenboom, J., and Kelly, T., Identifying, Understanding, and Analyzing Critical Infrastructure Independencies, *IEEE Control Systems Magazine*, 21, 11–25, 2001.

Shih, C. Y., Scown, C. D., Soibelman, L., Matthews, H. S., Jr, J. H. G., Dodrill, K., and Mcsurdy, S., Data Management for Geospatial Vulnerability Assessment of Interdependencies in U. S. Power Generation, *Journal of Infrastructure Systems*, 179–189, 2009.

Suguna, K., and Nandhini, K., Literature Review on Data Mining Techniques, *International Journal of Computer Technology and Applications*, 6, 583–585, 2015.

Wang, L., and Leite, F., *Knowledge Discovery of Spatial Conflict Resolution Philosophies in Bim-Enabled MEP Design Coordination Using Data Mining Techniques: A Proof-Of-Concept*, Computing in Civil Engineering, 2013.