Online Supplement to "Bibliometric Analysis and Critical Review of the Research on Big Data in the construction industry"

Normalization, Mapping, and Clustering technique of VOS

The keyword co-occurrence analysis is used to give a comprehensive review and research status quo. Firstly, VOSviewer (VOS) is used to analyze the literature information for mapping, clustering networks. The introduction of normalization, mapping, and clustering techniques (Van Eck and Waltman 2014) used by VOS is needed to present for better explanation and appropriateness.

There are considerable differences between nodes in the number of edges they have in a bibliometric network to other nodes. For instance, popular nodes representing highly cited publications or highly prolific researchers may have several orders of magnitude more connections than their less popular counterparts. In analyzing keyword co-occurrence networks in VOS, one usually performs a normalization for these differences between nodes. The Normalization method determines how the strength of the links between items is normalized. Normalized link strengths are used as input for the VOS layout technique and the VOS clustering technique. We use the association strength normalization (Van Eck and Waltman 2009) to normalize for differences between nodes in the number of edges they have to other nodes.

Let a_{ij} denote the weight (co-occurrences) of the edge between nodes i and j, where $a_{ij} = 0$ if there is no edge between the two nodes. Since VOS treats all networks as undirected, we always have $a_{ij} = a_{ji}$. The association strength normalization constructs a normalized network in which the weight of the edge between nodes i and j is given by

$$s_{ij} = \frac{2ma_{ij}}{k_i k_j} \tag{1}$$

where k_i (k_j) denotes the total weight of all edges of the node i (node j) and m denotes the total weight of all edges in the network. In mathematical terms,

$$k_i = \sum_j a_{ij} \tag{2}$$

$$m = \frac{1}{2}\sum_i k_i \tag{3}$$

The scholars sometimes refer to s_{ij} the similarity of nodes i and j. The similarity s_{ij} constructs the similarity matrix.

And the VOS mapping technique constructs a map based on the similarity matrix. The VOS mapping technique is used to position the nodes in the network in a two-dimensional space. The VOS mapping technique minimizes the function

$$V(\mathbf{x}_1,\cdots,\mathbf{x}_n)=\sum_{i<j}s_{ij}\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2 \tag{4}$$

subject to the constraint

$$\frac{2}{n(n-1)}\sum_{i<j}\left\|\mathbf{x}_i-\mathbf{x}_j\right\|=1 \tag{5}$$

where n denotes the number of nodes in a network, $\mathbf{x}_i$ denotes the location of the node i in a two-dimensional space, and $\left\|\mathbf{x}_i-\mathbf{x}_j\right\|$ denotes the Euclidean distances between nodes i and j. VOS uses a variant of the SMACOF algorithm (Groenen and Patrick 1997) to minimize (4) subject to (5).

Using the clustering technique of VOS, nodes are assigned to clusters by maximizing the function,

$$V(c_1,\cdots,c_n)=\sum_{i<j}\delta(c_i,c_j)(s_{ij}-\gamma) \tag{6}$$

where c_i denotes the cluster to which node i is assigned, $\delta(c_i,c_j)$ denotes a function that equals 1 if $c_i=c_j$ and 0 otherwise, and γ denotes a resolution parameter that determines the level of detail of the clustering. The higher the value γ, the larger the number of clusters that will be obtained. The function in (6) is a variant of the modularity function introduced by Newman and Girvan (2004)and Newman (2004) for clustering the nodes in a network. To validate the clustering techniques, VOS utilizes the smart local moving (SLM) algorithm in cluster analysis, whose performance, time complexity are better than other algorithms(Ludo Waltman 2013). The normalization, mapping, clustering techniques constitute a unified approach to mapping and clustering the nodes in keywork co-occurrence network.

The keyword co-occurrence analysis

The study of previous sections provides an evolutionary perspective for BD in the construction industry. With VOS's aid, we set the number of the minimum occurrences of a keyword to 10, and the normalized method is the Association strength normalization. So, of the 3707 keywords, 110 meet the threshold. For each of the 110 keywords. The keyword co-occurrence map indicates that the topics are divided into three clusters through the above cluster technique. The keyword co-occurrence analysis is applied to explore the logical architecture of BD in the construction industry. And we conclude some information on the typical papers related to each cluster, as shown in **Table 1**, **Table 2**, **Table 3**.

Table 1 Typical examples of Cluster 1 BD Application Scenario

Sub-category	Purpose of use	Opportunities or future work	Literature
Architecture sector	The parametric building design of many units	Automatic simulation and carbon analysis; Building a big parametric database	(Caetano and Leitao 2019); (Zhang et al. 2016)
Engineering sector	ML, GIS, decision-making methods in the	Engineering energy simulation and prediction	(Ngo 2019); (Antucheviciene et al. 2015); Sergi and Li (2014)

	civil engineering sector		
Construction sector	Classification of construction workers' mental fatigue	The application in the real construction site; Multi-level intervention strategies for mental fatigue.	(Li et al. 2020); Tekin and Atabay (2019)

Table 2 Typical examples of Cluster 2 BD Emerging Technology

Purpose of use	Model or algorithm	Opportunities or future work	Literature
Assess the building's environmental uncertainty	Fuzzy C-means clustering, the case-specific knowledge	Scenario uncertainty (e.g. assessment data source); Model uncertainty (e.g. building simulation program)	(Feng et al. 2019)
1.Detect Image-based object for site information retrieval and construction progress monitoring. 2. Reliable detection rates.	Classification Network, semantic segmentation, mask Region-based Convolutional Neural Network (R-CNN)	1.Detecting the construction elements by creating a CNN 2.A image-based construction monitoring process	(Braun and Borrmann 2019)
Predict litigation outcome of differing site condition disputes	SVM, NB, rule induction classifiers, DT, boosted decision trees (BDTs), the projective adaptive resonance theory (PART).	Other NLP in the construction disputes	(Mahfouz and Kandil 2012)
Classify object from the spatial and visual features	ANN, fuzzy logic (FL)	The intersection of BIM, visual and spatial sensing and sensor systems, computer vision, and image processing	(Brilakis et al. 2010)

Table 3 Typical examples of Cluster 3 BD Management

Sub-category	Purpose of use	Opportunities or future work	Literature
The diffusion of innovation theory	Promote BD diffusion	BD adoption rates/patterns with policy interventions; Facilitate BD policy development	(Gledson and Greenwood 2017); (Kassem and Succar 2017)

The barriers to digital innovation for BD	BD integration issues (formatting, privacy), IT tools for BD, skill requirement, BD management usability	A cultural change of trust; Data disclosure and privacy protection in a construction project	(Ahmed et al. 2018)

References

Ahmed, V., Aziz, Z., Tezel, A., Riaz, Z. (2018), “Challenges and drivers for data mining in the AEC sector”, *Engineering Construction and Architectural Management*, Vol. 25 No. 11, pp. 1436-1453.

Antucheviciene, J., Kala, Z., Marzouk, M., Vaidogas, E.R. (2015), “Solving civil engineering problems by means of fuzzy and stochastic MCDM methods: Current state and future research”, *Mathematical Problems in Engineering*, Vol. 2015, pp. 1-16.

Borg, I., Groenen, Patrick J.F. (1997), “Modern Multidimensional Scaling”, *Springer*.

Braun, A., Borrmann, A. (2019), “Combining inverse photogrammetry and BIM for automated labeling of construction site images for machine learning”, *Automation in Construction*, Vol. 106, pp. 1-12.

Brilakis, I., Manolis, L., Sacks, R., Savarese, S., Christodoulou, S., Teizer, J. (2010), “Toward automated generation of parametric BIMs based on hybrid video and laser scanning data”, *Advanced Engineering Informatics*, Vol. 24 No. 4, pp. 456-465.

Caetano, I., Leitao, A. (2019), “Integration of an algorithmic BIM approach in a traditional architecture studio”, *Journal of Computational Design and Engineering*, Vol. 6 No.3, pp. 327-336.

Feng, K.L., Lu, W.Z., Wang, Y.W. (2019), “Assessing environmental performance in early building design stage: An integrated parametric design and machine learning method”, *Sustainable Cities and Society*, Vol. 50, pp. 1-15.

Gledson, B.J., Greenwood, D. (2017), “The adoption of 4D BIM in the UK construction industry: an innovation diffusion approach”, *Engineering Construction and Architectural Management*, Vol. 24 No. 6, pp. 950-967.

Kassem, M., Succar, B. (2017), “Macro-BIM adoption: Conceptual structures”, *Automation in Construction*, Vol. 57, pp. 64-79.

Li, J., Li, H., Umer, W., Wang, H.W., Xing, X.J., Zhao, S.K., Hou, J. (2020), “Identification and classification of construction equipment operators' mental fatigue using wearable eye-tracking technology”, *Automation in Construction*, Vol. 109, pp.1-15.

Ludo Waltman, N.J.V.E. (2013), “A smart local moving algorithm for large-scale modularity-based community detection”, *European Physical Journal B*, Vol. 86 No. 11, pp. 471.

Mahfouz, T., Kandil, A. (2012), “Litigation outcome prediction of differing site condition disputes through machine learning models”, *Journal of Computing in Civil Engineering*, Vol. 26 No. 3, pp. 298-308.

Newman, M.E.J. (2004), “Fast algorithm for detecting community structure in networks”, *Physical Review E*, Vol. 69 No. 6, pp. 026113.

Newman, M.E.J., Girvan, M. (2004), “Finding and evaluating community structure in networks”, *Physical Review E*, Vol. 69 No. 2, pp. 026113.

Ngo, N.T. (2019), “Early predicting cooling loads for energy-efficient design in office buildings by machine learning", *Energy and Buildings*, Vol. 182, pp. 264-273.

Sergi, D.M., Li, J. (2014), "Applications of GIS-enhanced networks of engineering information", *Advances in computational modeling and simulation, Pts 1 and 2*, Vol. 444-445, pp. 1672-1679.

Tekin, H., Atabay, S. (2019), "Building information modelling roadmap strategy for Turkish construction sector", *Proceedings of the institution of civil engineers-municipal engineer*, Vol. 172 No. 3, pp. 145-156.

Van Eck, N.J., Waltman, L. (2014), "Visualizing bibliometric networks", *Measuring Scholarly Impact*, pp. 285-320.

Van Eck, N.J., Waltman, L. (2009), "How to normalize cooccurrence data? An analysis of some well-known similarity measures", *Journal of the American Society for Information Science and Technology*, Vol. 60 No.8, pp. 1635-1651.

Zhang, Y.A., Zhu, Z.Y., Li, C.G., Chang, L. (2016), "Integration application system of Chinese wooden architecture heritages based on BIM", *2016 international conference on logistics, informatics and service sciences*, New York: Ieee.