JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

Research Paper
J. of infosci. theory and practice 2(2): 16-32, 2014
http://dx.doi.org/10.1633/JISTaP.2014.2.2.2

Exploratory Study of Developing a Synchronization-Based Approach for Multi-step Discovery of Knowledge Structures

So Young YU *

Department of Library and Information Science
Hannam University, Republic of Korea
E-mail: soyoungyu201@gmail.com

ABSTRACT

As Topic Modeling has been applied in increasingly various domains, the difficulty in naming and characterizing topics also has been recognized more. This study, therefore, explores an approach of combining text mining with network analysis in a multi-step approach. The concept of synchronization was applied to re-assign the top author keywords in more than one topic category, in order to improve the visibility of the topic-author keyword network, and to increase the topical cohesion in each topic. The suggested approach was applied using 16,548 articles with 2,881 unique author keywords in construction and building engineering indexed by KSCI. As a result, it was revealed that the combined approach could improve both the visibility of the topic-author keyword map and topical cohesion in most of the detected topic categories. There should be more cases of applying the approach in various domains for generalization and advancement of the approach. Also, more sophisticated evaluation methods should also be necessary to develop the suggested approach.

Keywords: Synchronization, Ego-centric Network, Topic Modeling, Informetrics

1. INTRODUCTION

The approach of Topic Modeling ("TM") has been applied in various domains, such as tech forecasting, text mining, and informetrics (Griffiths & Steyvers, 2004; Kang et al., 2013; Lu & Zhai, 2008; Park & Song,

Open Access

Accepted date: June 24, 2014
Received date: June 4, 2014

***Corresponding Author:** So Young YU
Assistant Professor
Department of Library and Information Science
Hannam University, Republic of Korea
E-mail: soyoungyu201@gmail.com

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (http:/ creativecommons.org/licenses/by/3.0/). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

ⓒ So Young YU , 2014

2013; Song et al., 2013; Tang et al., 2012; Titov & McDonald, 2008; Yu, 2013). The foundation and applicability of TM can be described as being fundamental and concrete due to the fact that is based on the Probability Model, which is linked to the Language Model. Development of an advanced method, at the same time, has been an ongoing process in order to enhance the performance of TM and various tools, and its applications have been developed and distributed.

The approach to interpreting the result of TM, however, has been recognized as an area needing growth in resolving the difficulty in characterizing and interpreting topics. Evidential cases of this difficulty have been found in some studies and these studies demonstrate additional efforts, such as inserting topics or eliminating keywords, to reduce the difficulty of interpretation (Hall et al., 2008; Talley et al., 2011). For example, non-informative keywords were eliminated and more informative phrases were added for easier interpretation after executing LDA-based TM analysis on NIH-supported research output in the study of Hall et al. (2008). Similar to the study of Hall et al. (2008), some number of topics were inserted additionally after finding 36 valid topics by applying TM in the research of Talley et al. (2012).

One of the possible reasons for this difficulty could be the de-contextualization of the relation among the top-k keyword in a topic when the output of TM is provided. The most common way of providing the output of TM is a list of top-k keywords of each topic (Chang et al., 2009). The sorted keywords in order of probability, however, would not be enough to deliver the character of a certain topic, which could be inferred from the overall combination of the loaded keywords in the topic (Chuang et al., 2012; Ramage et al., 2009b). And this implies that characterizing the latent topic could be needed in additional works for inferring meaningful contexts from the keyword list.

Another possible reason could be a multi-assigned keyword; that is, a keyword which is assigned to more than two topics at the same time with high probability. This means that the multi-assigned keywords are, probably, frequently occurring keywords in a certain dataset and this could make interpretation vague and lead to several similar names among the topics. Therefore, it could be hard to achieve the distinctiveness of interpretation.

Sophisticated methods of TM and visualization have been suggested to make interpretation easier (Chaney & Blei, 2012; Chuang et al., 2012). Labeled LDA (Ramage et al., 2009a) and Partially Labeled LDA (Ramage et al., 2011) were developed for enhancing the performance of TM. Several visualization approach and network analysis methods including citation linking have been also applied for the better performance of TM (Mei et al., 2008; Nallapati et al., 2008).

Along with the previous research, this study, therefore, aims to explore an approach to enhance the ease of interpretation by combining social network analysis with topic modeling. In order to contextualize the keywords in a topic and to reduce the number of multi-assigned keywords, co-word analysis, and the concept of a "synchronization network" is applied in refining the result of TM without distracting the topical cohesion in a topic.

Synchronization in a complex network is defined as phase transition when the entire network of nodes begins to emit and receive a signal at the same frequency, and this phenomenon has been detected and researched in various domains (Arenas et al. 2008; Kuramoto & Nishikawa , 1987; Niebur et al., 1991; Pikovsky et al., 2001; Strogatz, 2000; Strogatz, 2001; Strogatz, 2003; Strogatz & Mirollo, 1988). Synchronization can be understood as a dynamic of networks focusing on the change of the property of a node affected by the property of the group of its connected nodes. Various applications of synchronization, such as analysis of genetic networks, systemic analysis on neuronal networks, data mining, opinion dynamics, neuroscience, or social sciences have been developed (Blasius et al. 1999; Buchanan, 2007; Elowitz & Leibler, 2000; Garcia-Ojalvo et al., 2004; Pluchino et al., 2005).

Applications of synchronization in data mining have been proposed for data clustering. The assumption for applying synchronization on data mining is that the dynamics of the data system could be categorized into clusters by detecting synchronization, and most of the previous research focused on dynamic modeling for the detection. Based on statistical methods of data mining, therefore, synchronization has been used in sophisticating the data mining techniques and exploiting the applicability of synchronization in data mining (Jalili, 2013; Jha & Yadava, 2012; Miyano & Tsutsui, 2007a; Miyano & Tsutsui, 2007b; Miyano & Tsutsui, 2008a; Miyano & Tsutsui, 2008b; Miyano & Tsutsui, 2009; Miyano & Tsutsui, 2013;Tilles et al., 2013; Wan et al., 2010).

Along with the previous studies, the application of

synchronization in text mining was explored in this study by applying the concept in deliberating the result of topic modeling. There are operational definitions for the application. It is assumed that a multi-assigned keyword can be re-assigned to a certain topic by synchronizing the topic of the keyword with those of its co-occurring keywords. In this study, therefore, "keywords in co-word network" is matched to "the nodes" in the synchronization network and "re-assignment of one topic to a multi-assigned keyword" is matched for "phase transition." The "entire network" is defined operationally as an ego-centric network of a certain multi-assigned keyword and its connected keywords that were used for determining the topic for the ego.

2. METHODOLOGY

2.1. Research Design

This study suggested the combined approach of text mining and network analysis. The research design of this study is shown in Figure 1. The perplexity from the natural language processing domain was considered in the first step, and topical similarity from text mining was applied in LDA (Latent Dirichlet Allocation, Blei et al., 2003)-based Topic Modeling ("LDATM") and Merging Overlapped Topics ("MOT") steps. The concept of synchronization networks and ego-centric networks from complex networks was applied in the Re-Assigning Multi-Assigned Keyword ("RAMAK") step.

In the subject of Construction & Building Technology, 16,584 bibliographic records of KSCI-indexed articles[1] were collected and pre-processed for topic modeling and co-word analysis. The indexed keywords for the analysis were from an English authors' keyword field, and 2,881 keywords were used in the modeling.

LDATM was performed after 10 times of pre-testing for finding the optimal number of topics, and the top 20 keywords for each topic were selected. After modeling, all pairs of topics with similarity values of 1 by comparing all the probability of the loaded keywords were merged as one topic. The multi-assigned keywords were also identified after merging topics.

For the re-assigning process (RAMAK), a co-word network of 2,486 top 20 keywords was extracted by calculating cosine similarity between keywords from a document-keyword matrix. The Ego-centric network of each multi-assigned keyword was extracted and re-assigned a topic on the ego that was determined by finding the most frequently occurring topic number from its nearest neighbor keywords, with the cosine similarity of their connections in mind.

The details of each process are as follows in 2.2, 2.3, 2.4 and 2.5.

2.2. Data Collection and Pre-Processing

To begin, 16,548 KSCI-indexed articles in the domain of construction and building technology were collected for the analysis. The data fields for articles collected were: publication year, DOI, ISSN, journal name, citation counts, title, author name, affiliation,

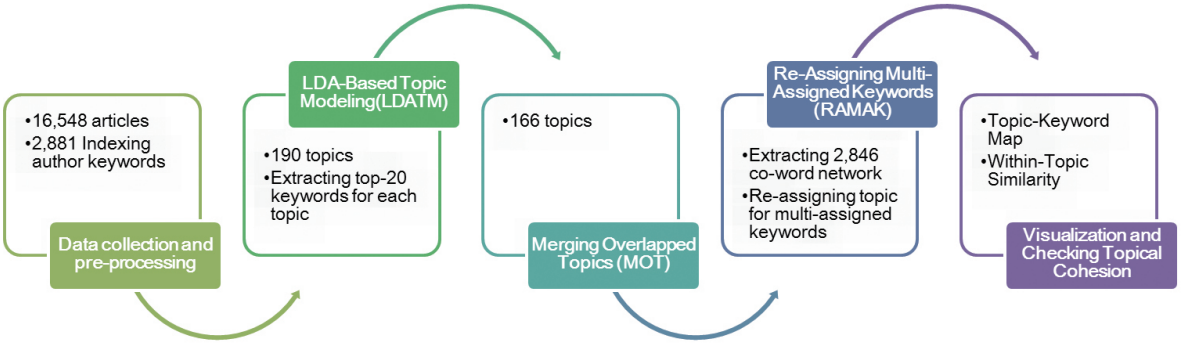


Fig. 1 Research design

[1] KSCI (Korea Science Citation Index) is one of the national citation indices of the Republic of Korea and it provides bibliographic records of articles published in 661 national major journals of science and technology. See http://ksci.kisti.re.kr/main/about.ksci.

author keywords, and abstracts. All the text fields, such as title, author keywords, or abstracts were written both in English and Korean. For this study, author keywords were indexed and all syntactic stopwords were eliminated. The total number of indexed terms from author keyword fields is 2,881 as shown in Table 1.

Table 1. Data Collection

KSCI Subject category	No. of indexed journals	No. of indexed articles	No. of citing articles	No. of references	No. of indexed author keywords
Construction & Building Technology	36	16,548	12,425	234,849	2,881

In order to set the number of topics, a perplexity score was calculated with 1,030 times of LDA-Based Topic Modeling. Perplexity score in a corpus means the predictability of a topic model and it is a widely-used metric in topic model evaluation (Asuncion et al., 2009).

Perplexity score is a parameter of how well a probability model predicts a test set (sample) in information theory and measurement of evaluating Language Modelin natural language processing. A lower value means a surer model (Brown et al., 1992).

The modelings were executed with different numbers of topics and 1,000 iterations, and the number of topics in a model has been changed in the range from 10 to 1030 by increasing 10 for each modeling. The number of topics with the lowest perplexity score was estimated as 190.

Stanford Topic Modeling Toolbox 0.4.0 ("TMT," Ramage et al., 2009a) was used in pre-testing. TMT, which has been developed by Stanford National Language Lab, was aimed to support research in social sciences and related fields by applying topic modeling on textual data. It is based on Java and supports LDA, Labeled LDA, and Partially Labeled LDA (Blei et al., 2003; Blei et al., 2006; Ramage et al., 2011).

As a result, the optimal number of topics was estimated as 190 by using perplexity value. The most frequently identified number of topics with a minimum value of perplexity was 190.

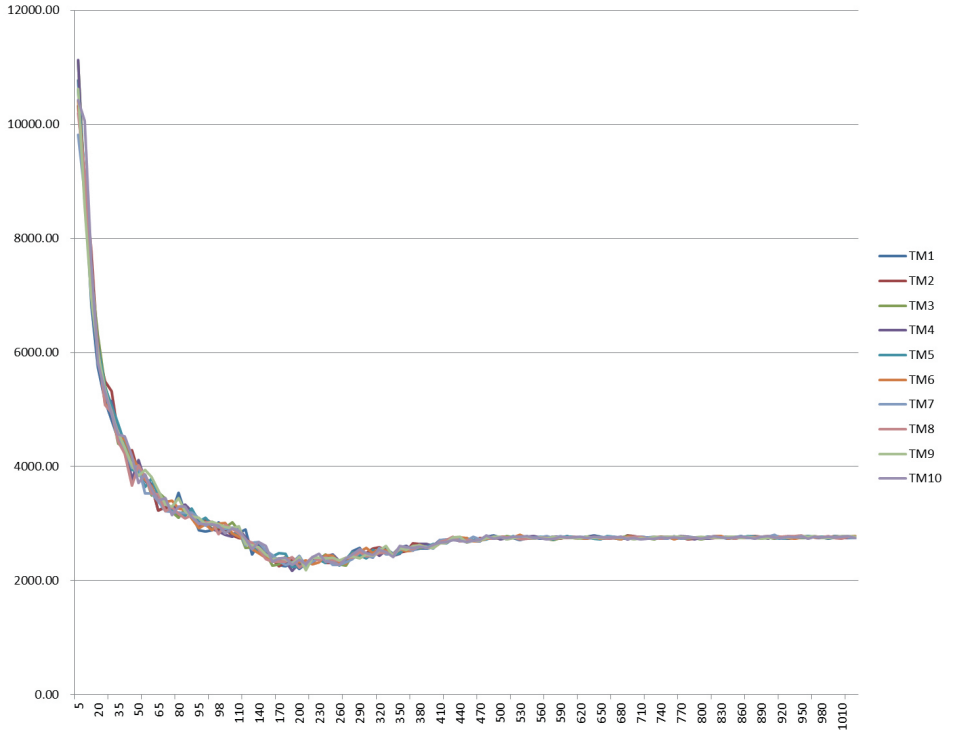


Fig. 2 Changes in perplexity value over the number of topic

Table 2. Perplexity changes in pre-test

# of Topic	LDATM1	LDATM2	LDATM3	LDATM4	LDATM5	LDATM6	LDATM7	LDATM8	LDATM9	LDATM10
10	8975.335	8963.464	8831.376	8649.377	9084.956	9109.842	8783.968	9147.761	8550.985	10055.105
100	2816.215	2792.832	3026.406	2769.778	2824.944	2828.139	2931.039	2904.674	2915.727	2923.103
110	2838.844	2741.301	2868.584	2911.134	2802.437	2768.915	2884.796	2821.816	2957.938	2876.513
120	2899.212	2733.605	2570.985	2759.323	2662.365	2696.362	2616.600	2713.299	2643.197	2733.443
130	2462.717	2606.532	2586.339	2632.329	2637.865	2551.660	2665.585	2666.303	2624.515	2547.768
140	2629.703	2537.296	2529.983	2598.290	2532.595	2481.425	2676.373	2545.855	2584.815	2675.385
150	2543.644	2529.799	2481.421	2493.457	2450.562	2398.660	2526.935	2375.752	2454.982	2613.446
160	2363.844	2378.645	2261.015	2358.277	2424.803	2345.375	2452.961	2334.453	2384.007	2332.292
170	2280.889	2251.022	2281.229	2362.431	2480.557	2386.532	2339.762	2334.752	2392.494	2392.639
180	2254.074	2326.227	2426.930	2339.651	2467.936	2378.713	2280.134	2337.829	2376.632	2401.485
190	2311.001	2366.047	2301.056	2176.891	2236.664	2405.793	2309.904	2417.616	2286.746	2270.426
200	2210.481	2317.458	2313.562	2309.178	2249.580	2323.953	2431.527	2223.865	2408.394	2368.890
210	2296.630	2342.619	2267.309	2291.714	2263.767	2351.669	2257.607	2312.559	2182.767	2286.754
220	2380.728	2321.671	2370.908	2392.573	2414.567	2285.605	2327.877	2415.950	2397.257	2412.482
230	2387.080	2402.662	2362.987	2405.343	2369.918	2324.615	2402.163	2351.932	2400.145	2469.383
240	2452.924	2314.648	2455.086	2425.034	2313.190	2452.122	2403.438	2410.961	2387.761	2333.303
250	2304.188	2315.152	2346.487	2455.857	2445.545	2436.064	2274.259	2351.760	2405.131	2319.241
300	2399.035	2487.912	2414.451	2460.291	2386.524	2578.468	2442.475	2429.788	2457.273	2480.635
400	2583.301	2575.829	2585.134	2603.352	2644.606	2579.357	2561.219	2590.439	2572.175	2634.949
500	2758.263	2769.149	2735.997	2720.636	2761.117	2770.918	2768.335	2758.400	2765.126	2758.162
600	2780.201	2766.005	2745.685	2746.498	2774.380	2755.659	2745.235	2752.476	2756.612	2759.085
700	2736.222	2781.246	2766.452	2761.840	2747.509	2764.499	2773.267	2766.780	2725.437	2741.442
800	2751.469	2733.360	2727.797	2752.880	2743.352	2748.101	2759.525	2756.007	2771.173	2740.718
900	2755.025	2768.088	2762.950	2735.265	2747.222	2777.986	2771.064	2752.937	2749.728	2782.771
1000	2762.954	2776.877	2744.741	2762.338	2780.410	2771.742	2769.924	2751.173	2752.731	2757.215
1010	2760.813	2772.226	2746.074	2743.839	2769.542	2774.740	2756.311	2734.745	2781.271	2765.664
1020	2744.803	2760.696	2772.429	2745.061	2762.559	2774.252	2755.717	2771.226	2782.883	2777.400
1030	2771.209	2757.181	2778.056	2768.634	2745.139	2782.631	2745.651	2765.064	2765.616	2745.922
No_Topic	200	170	160	190	190	220	210	200	210	190

Note: The rest of this table is available upon request.

2.3. Topic Detection and Merging Similar Topics

By setting the number of topics as 190, LDA-based Topic Modeling (LDATM) was performed with 1,000 repetitions. Topic Modeling Toolbox 0.4 was used for the modeling. Each topic was labeled with its topic number, such as 'T1,' and the most dominant topic had a smaller number in the label. The top 20 keywords of each topic were selected to describe the topic and the total number of unique top keywords was 2,846. Stanford TMT 0.4.0 was also used in topic detection.

Merging overlapped Topics (MOT) was executed by calculating similarity between two topics, and the topics are merged at a threshold of 1. The merged topic that 25 topics were merged into was labeled as "T4."

2.4. Extracting Co-Word Network and Re-arranging Multi-Assigned Top-Keywords

The purpose of this step was to re-arrange a multi-assigned top-keyword in LDA-Based TM by assigning the keyword into one topic. There were 609 top-keywords which were assigned to more than two topics at the same time.

In this step, the concept of oscillator in synchronization networks was applied practically to determine and assign one topic to a multi-assigned keyword. In this study, the suggested application is more focused on the result of synchronization itself as a change of the property of a node evoked by its neighbor nodes, rather than modeling the dynamics of synchronization, in order to use the concept of synchronization to assign one topic to a multi-assigned keyword considering the topics which its nearest neighbors were assigned to. That is, the topic of the multi-assigned keyword is determined by the most frequently occurring topic from its nearest neighbors.

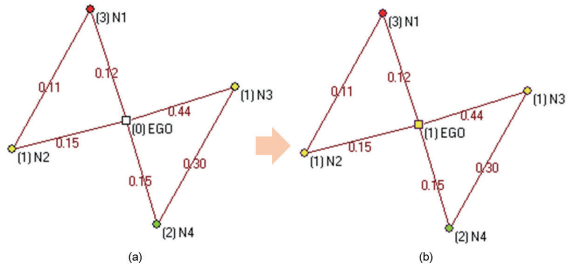
The process of applying synchronization is:

a. Extract ego-centric network of a multi-assigned keyword from co-word network.
b. For each neighbor, check the assigned topic numbers.
c. For each topic number pertinent to the neighbor, sum the cosine similarity between ego and its neighbor.
d. Iterate "b-c" for all neighbor nodes.
e. Average the summed cosine similarity for each topic number.
f. Determine the most influential topic number that has the maximum value of averaged cosine similarity.

To extract an ego-centric network, co-occurrence networks among 2,846 top-keywords were extracted by using cosine similarity. For 609 multi-assigned keywords, 609 individual ego-centric networks were extracted. In each network of "EGO," the ego node is a certain multi-assigned keyword. The network consists of the ego and its nearest neighbors, which are directly connected to the ego. The link weight is the cosine similarity between two nodes.

When each keyword is assigned to a certain topic as a result of LDATM and the links are calculated by using cosine similarity, the combination of those LDATM and co-occurrence similarities is used to determine the topic for the "EGO."

For example, assume that there is an ego-centric network for a multi-assigned keyword "EGO" with four nearest neighbor keywords, "N1," "N2," "N3," and "N4." This is a subset of the co-word network consisting of "EGO" and its nearest neighbors as node and links among them. The link weight between two nodes is the cosine similarity score of their co-occurrence in a document. To determine the topic category of the "EGO," the weight score for a certain topic is averaged by using the cosine similarity between the "EGO" and its connected keywords that are assigned in the topic. For instance, N1 and N3 were assigned to Topic [1], while N4 for Topic [3], and N2 for Topic[2] were assigned, respectively, in Figure 3. The topic for "EGO" was re-assigned as Topic[1] because the highest weight was from the combination of N2 and N3 with Topic[1], as shown in Figure 3-(b).



Topic number	Neighbor	Avg. cosine similarity
[1]	N2, N3	(0.15+0.44)/2 = 0.295
[2]	N4	0.15
[3]	N1	0.12

Fig. 3 Example of RAMAK: Re-assigning a topic to a multi-assigned keyword

2.5. Comparison of Maps and Topical Cohesion

To identify the effect of applying RAMAK, comparison of topic-keywords networks was conducted. The topic-keywords maps were made by using PAJEK 3.0 for visual comparison. For seeing a structural difference, the density and average degree of nodes were calculated and compared.

For comparing topical cohesion, "within-topic cosine similarity" was computed. The value of within-topic cosine similarity is an average cosine similarity of all occurring pairs of keywords in a specific topic. For a topic i, cosine similarity (i) for all pairs of keyword$_j$ and keyword$_k$ in the topic i were summed up and divided by the number of the pairs.

Within-topic similarity $(i) = \frac{1}{n}\sum_{i=1}^{n} cosine_i(kwd_j, kwd_k)$

The differences in within topic similarity for all topics were statistically tested by conducting a paired-sample t-test and correlation analysis using SPSS21.

3. RESULTS AND ANALYSIS

3.1. Changes in Major Topics and Related Keywords

Each topic has top 20 keywords after applying LDATM and the number of the top keywords in several topic categories was changed by applying MOT and RAMAK. The number of 20 top keywords in only three topics (T157, T57, and T83) had been kept after reassigning the multi-assigned keywords, and the number of top-keywords in the other 163 topics had been changed. Only three topic categories (T15, T42, and T74) were expanded in terms of size, and another 160 topic categories resulted in fewer top-keywords. This doesn't mean the decrease of the number of top-keywords but the decrease of the number of multi-assigned keywords.

Figure 4 –(a) shows the number of changes in the number of top-keywords for each topic category and Figure 4 –(b) shows the number of topic categories with the size of the topic. The sizes of topics mostly decreased, and only three topics were increased in size. The topics which resulted in less than 10 top keywords are T79 and T612, while the topics with more than 20 top keywords are T15, T42, and T74.

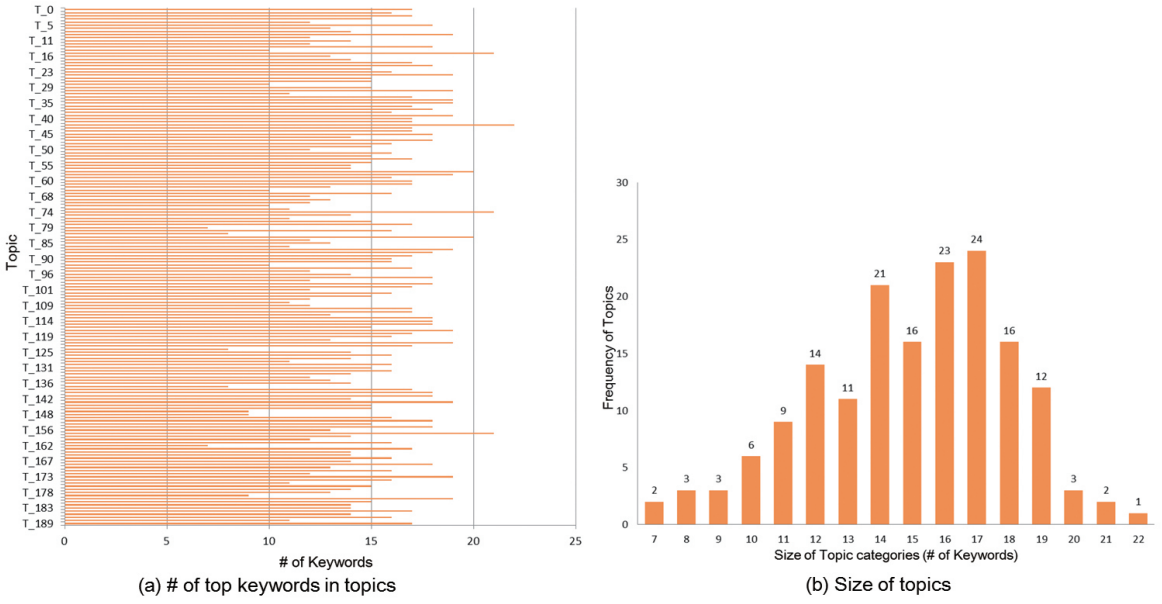


Fig. 4 Changes in the size of topics

The number of keywords being assigned in more than two topics ("multi-assigned keyword") was 609, which is 21.3% of the total number of unique top keywords (2,864). The number of multi-assigned top-keywords was also decreased to 7, and 602 out of 609 multi-assigned keyword were re-assigned to one topic category by applying RAMAK. The remains of multi-assigned keywords were "Shape memory alloy," "Urban Residential Area," "Design Value Engineering," "Peter Zumthor," "Brittle fracture," "Fatigue crack," and "Housing Satisfaction," which cover several research topics of construction and building technology in general.

The top 10 topics with loaded top keywords were shown in Table 4. As shown in the table, multi-assigned keywords were re-assigned to non-top 10 topics and new top keywords from non-top 10 topics were added to any of the top 10 topics.

For example, the word term of "optimal design" in italics had been loaded to three topics among the top 10 topics; that is, T2, T4, and T8 in LDATM result, and it had been loaded to 33 different topics among the 190 topics. It was re-loaded to T100 after applying MOT and RAMAK. And the word term of "Mixed-Use Development" in T3 and T5 was re-located in T15 by applying MOT and RAMAK.

Adding new top keywords to the top 10 topic categories was made by the application. For example, "global warming" (underlined) in T3 was added to T3 by considering the most prominent topic number among its strongly connected co-words. All newly added top-keywords in the top 10 topics were underlined in Table 4.

Table 3. Major Multi-Assigned Keywords

Keywords	LDATM+MOT	LDATM+MOT+RAMAK	LDATM
Radar rainfall	9	1	33
Optimal design	9	1	33
Forest fire	9	1	33
Conversion	7	1	31
Flexural strength	7	1	7
Sediment transport	6	1	30
Vegetation	6	1	30
Furniture design	6	1	30
Productivity	6	1	6
Composite	6	1	6

Table 4. Comparison of Top 10 Topic Categories and Its Keywords

Topic	20 Top-Keywords (LDATM)	Top-Keywords (LDATM+MOTRAMAK)
T0	Shaking Table Test, Partial Safety Factor, Nuclear Power Plant, Science Museum, Slope Stability, Cold-Formed Steel, Limit State Design, Displacement Ductility, Exhibition Method, Anchor Bolt, EDG(Emergency Diesel Generator), Near-Fault Ground Motion, Flexural Capacity, Reliability-Based Design, Holding Power, Children, Seismic Capacity, RC building, Safety Factor, Sliding	Shaking Table Test, Partial Safety Factor, Nuclear Power Plant, Science Museum, Slope Stability, Cold-Formed Steel, Limit State Design, Exhibition Method, Anchor Bolt, Near-Fault Ground Motion, Reliability-Based Design, Holding Power, Children, RC Building, Safety Factor, Sliding, EDG(Emergency Diesel Generator)

T1	Permeability, Infiltration, Urbanization, Hysteresis, SWMM, Sustainable Design, Soil-Water Characteristic Curve, Construction Project, Unsaturated Soils, Competency, Competency Model, Green Home, City, Sensor, Unsaturated Soil, Natural Elements, Redevelopment, Groundwater Level, Mix Design, Insulation	Permeability, Infiltration, Urbanization, Hysteresis, Sustainable Design, Soil-Water Characteristic Curve, Construction Project, Unsaturated Soils, Competency, Competency Model, Green Home, City, Unsaturated Soil, Natural Elements, Mix Design, ESP-r
T2	Small Hydropower, Methanehydrate, Natural gas, Design Parameter, Storage, Specific Output, Rainfall Condition, Equilibrium, Design Flowrate, Load Factor, Trench, Heliostat, Direct Normal Insolation, Capacity, Receiver, Cooling Energy, Flow Duration Curve, Priority, Lighting Energy, *Optimal Design*	Small Hydropower, Methanehydrate, Natural gas, Design Parameter, Storage, Specific Output, Rainfall Condition, Equilibrium, Design Flowrate, Load Factor, Trench, Heliostat, Direct Normal Insolation, Capacity, Receiver, Cooling Energy, Correlation
T3	Urban Design, Housing, Urban Landscape, Correlation Analysis, Mixed-Use Development, Urban Spatial Structure, Development, Sea Surface Temperature, Streetscape, Urban, CBD, Integration, Meaning, Residential Environment, Cognitive Map, Design Guidelines, Plan, Architecture, Sense, Circulation System	Urban Design, housing, Urban Landscape, Urban Spatial Structure, Development, Sea Surface Temperature, Urban, CBD, Integration, Meaning, Cognitive Map, Plan, Sense, Circulation System, Global Warming
T4	De Stijl, VRS, Optimal Design, Forest Fire, Radar Rainfall, Sediment Transport, Conversion, Kalman Filter, Regional Characteristics, Pan Evaporation, Space-Time, MCDM, Reference Evapotranspiration, Transition, Furniture Design, CORS, Vegetation, Multi Criteria Decision Making, Neo Plasticism, Work Information	De Stijl, VRS, Kalman Filter, Regional Characteristics, Pan Evaporation, Space-Time, MCDM, Reference Evapotranspiration, CORS, Multi Criteria Decision Making, Neo Plasticism, Work Information
T5	Land Use, Model, Thermal Environment, Public Design, Urban Planning, Urban Climate, Shallow-Water Equations, k- Varepsilon, Surface Roughness, RANS, Finite Volume Method, Street Furniture, Landscape, Kappa, Dam-Break, Heat Island, Urban Climate Simulation System, *Mixed-Use Development*, Approximate Riemann Solver, Urban Heat Island	Land Use, Model, Public Design, Urban Planning, Urban Climate, Shallow-Water Equations, k- Varepsilon, Surface Roughness, RANS, Finite Volume Method, Street Furniture, Kappa, Dam-Break, Urban Climate Simulation System, Approximate Riemann Solver, Urban Heat Island, Method, Vortex
T6	Drying Shrinkage, Autogenous Shrinkage, Diffusion Coefficient, Chloride Penetration, Bridge Deck, Early Age, Finishing Material, Chloride Ion, Water Content, Architectural Space, Porosity, Diffusion, Cracking, Regression Analysis, Steel Powder, Convection, Fiber, Hydration Heat, Mechanical Behavior, Humidity	Drying Shrinkage, Autogenous Shrinkage, Diffusion Coefficient, Chloride Penetration, Bridge Deck, Early Age, Finishing Material, Chloride Ion, Cracking, Steel Powder, Convection, Hydration Heat, Mechanical Behavior
T7	FEM, Beam-To-Column Connection, Elastic Modulus, Diaphragm, Stress Concentration, Soft Ground, Consolidation, Shear Buckling, Anisotropy, Steel Box-Girder, Stiffener, Parametric Study, Cyclic Loading Test, Strength, Embankment, Steel Box-Girder Bridge, RBDO, Connection, Reliability Based Design Optimization, Added Mass	FEM, Beam-To-Column Connection, Diaphragm, Stress Concentration, Soft Ground, Shear Buckling, Anisotropy, Steel Box-Girder, Stiffener, Steel Box-Girder Bridge, RBDO, Reliability Based Design Optimization, Added Mass, Fatigue Crack
T8	De Stijl, VRS, *Optimal Design*, Forest Fire, Radar Rainfall, Sediment Transport, Conversion, Kalman Filter, Regional Characteristics, Pan Evaporation, Space-Time, MCDM, Reference Evapotranspiration, Transition, Furniture Design, CORS, Vegetation, Multi Criteria Decision Making, Neo Plasticism, Work Information	Being merged with T4
T9	Semi-Rigid Connection, Steel Frame, Plastic Hinge, Beam-Column, Structural Optimization, Semi-Rigid, Design Factor, Diagrid, Push-Over, Dynamic Relaxation Method, Shape Optimization, Initial Stiffness, Story Drift, Arc-Length Method, Elasto-Plastic Analysis, Theta, Pushover Analysis, Non-Linearity, Curve, Architectural Planning	Semi-Rigid Connection, Steel Frame, Plastic Hinge, Structural Optimization, Semi-Rigid, Design Factor, Diagrid, Push-Over, Dynamic Relaxation Method, Shape Optimization, Story Drift, Arc-Length Method, Elasto-Plastic Analysis, Theta, Non-Linearity, Curve, Arrangement, Brittle Fracture

Note. The rest of this table is available upon request.

3.2. Changes in Knowledge Structure

Both the density and average degree of the topic keyword network were decreased by applying MOT+RAMAK, due to the fact that the links between multiple topics and top keywords were eliminated. Therefore, the network of "LDATM+MOT+RAMAK" was identified as providing a clearer view visually.

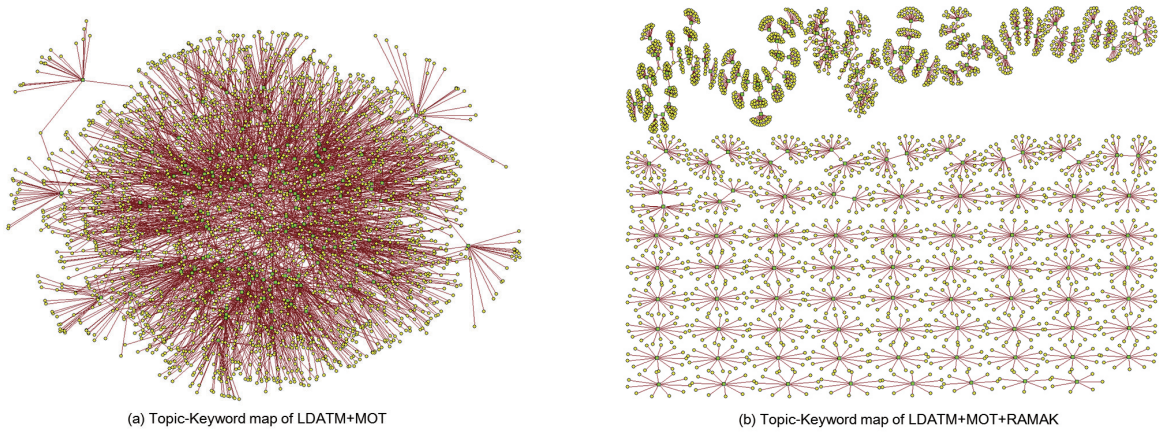
The enhancement in visualization could be identified in network structure in terms of density and average degree. The density of the topic-keyword network of LDATM+MOT+RAMAK was decreased to 0.006 from 0.008, so that it could be expected to have less complex link structures. The average number of connected nodes for each node in the network (Avg. Degree) was also decreased to 1.88 from 2.84.

The improvement in visualization was also presented in comparing those maps visually. As shown in Figure 5–(a) and Figure 5 –(b), it can be visually identified that the "LDATM+MOT+RAMAK" network has a more refined and clearer structure between topics and their top keywords. The box indicates a topic category and the eclipse indicates a certain top keyword.

There was only one component in Figure 5-(a), in which all topics and keywords were connected by the multi-assigned keywords. It was obvious that all topics in the analyzed domain were related to each other, but interpretation was difficult because of the visual complexity. The map of Figure 5-(b), which was the result of applying RAMAK, presented a relatively clearer and more easy-to-analyze situation. The number of connected topics and keywords was decreased and most of the topics were differentiated from each other visually.

Table 5. Structure of Topic-Keyword Networks

KSCI Subject category	# of Top-20 Keyword nodes	# of Topic nodes	Density	Avg. Degree
LDATM	2,486	190	0.008	2.84
LDATM+MOT	2,486	166	0.008	2.5
LDATM+MOT+RAMAK	2,486	166	0.006	1.88



(a) Topic-Keyword map of LDATM+MOT

(b) Topic-Keyword map of LDATM+MOT+RAMAK

Fig. 5 Changes in the structure of topic-keyword maps

In addition to the change in overall topic-keywords network structure, a change in the 10 dominant topics and their pertinent keywords was also analyzed. The number of keywords in Figure 6 was 178 and the number of keywords in Figure 7 is 141, because of re-assigning the multi-assigned keywords. As was identified in Table 4, two multi-assigned keywords, that is, "optimal design" and "multi-use development" were re-located to other topic categories and links between T3 and T5 and between T2 and T4 were eliminated.

It was identified that the changes in dominant topics were not huge, relatively, when comparing the result to Figure 5. It could, therefore, imply that the process of re-assigning multi-assigned keywords could keep the major knowledge structure while refining the relatively peripheral topics.

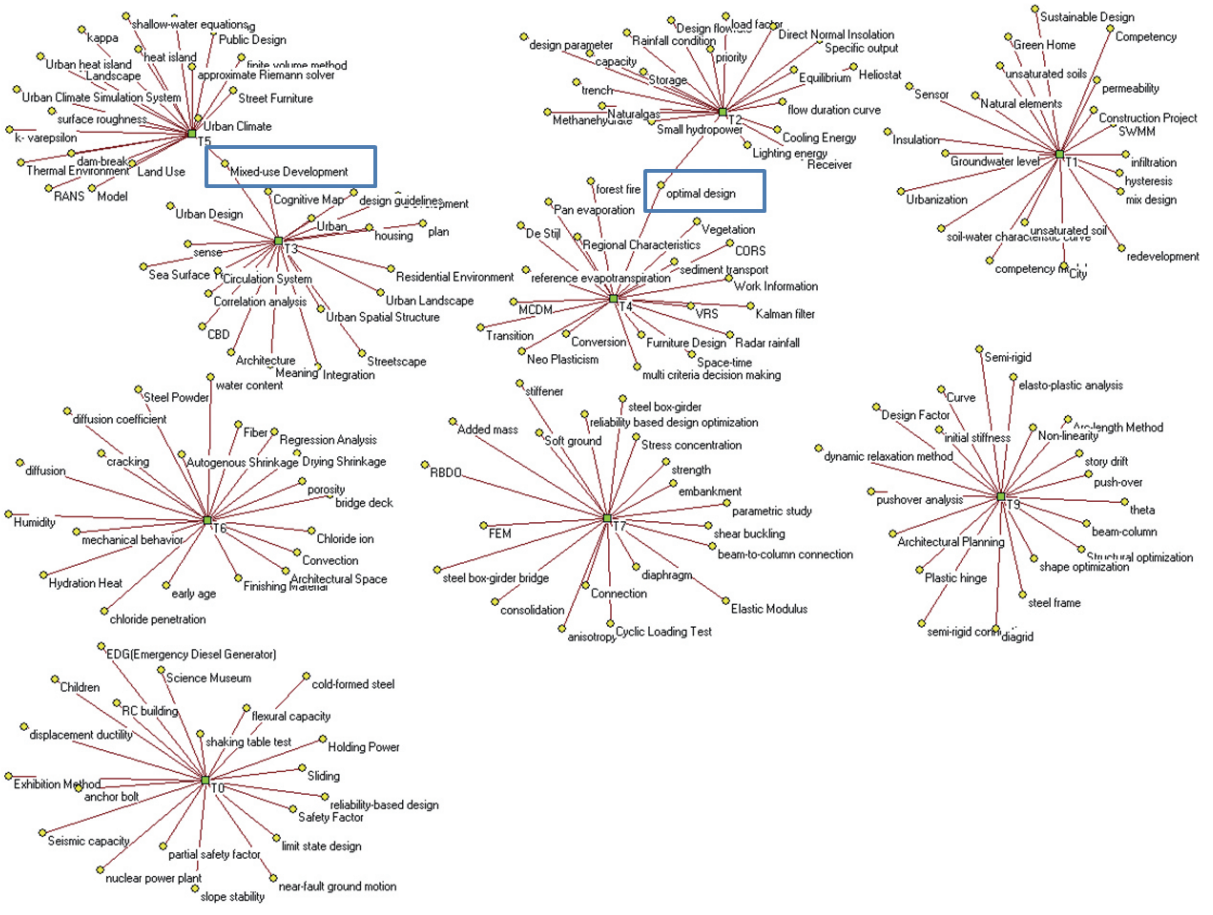


Fig. 6 Dominant topic-keyword map of LDATM+MOT

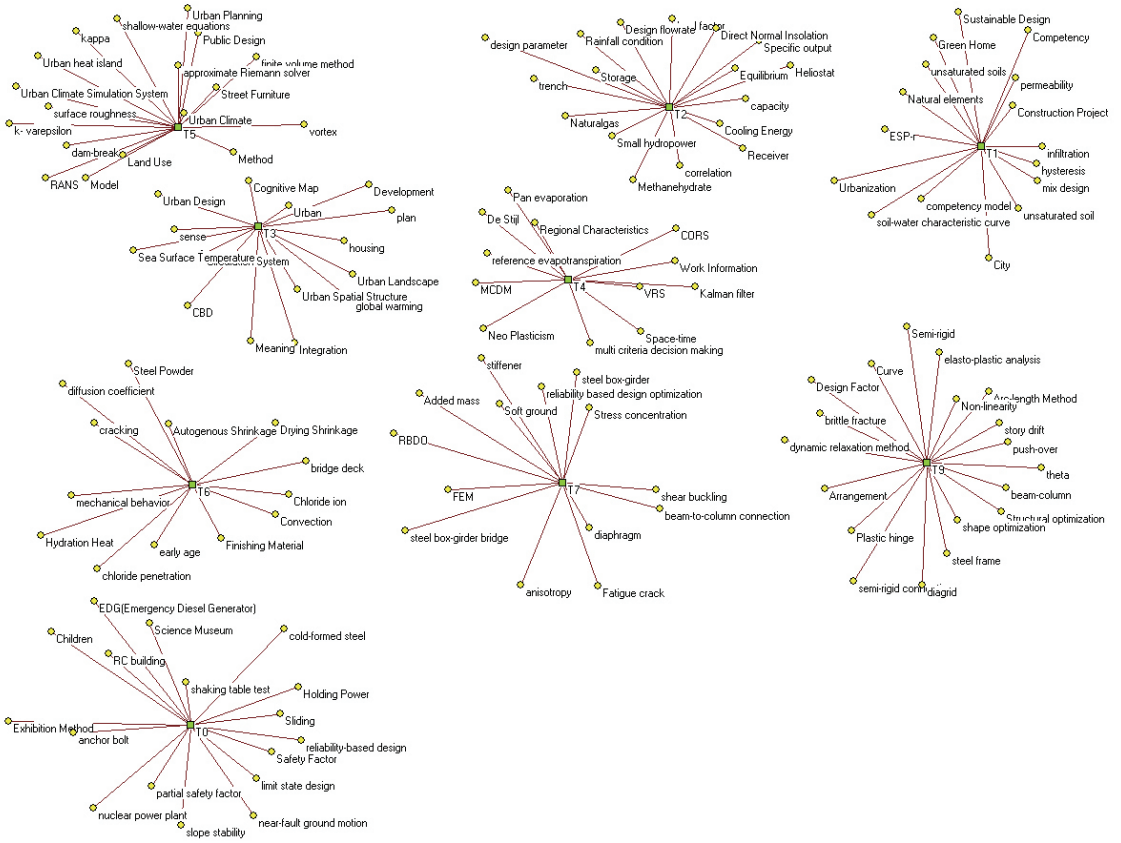


Fig. 7 Dominant topic-keyword map of LDATM+MOT+RAMAK

3.3. Changes in Topical Cohesion

The average value of within-topic cosine similarity over 166 topic categories was increased by 0.029 after re-assigning the multi-assigned keywords, and the difference was statistically significant ($t(166) = -16.27$, $p < .01$). The values of within-topic similarity of 159 topic categories were increased and only seven topic categories had decreased within-topic cosine similarity values.

A paired-samples t-test, therefore, was conducted to evaluate whether the within-topic similarity of LDATM+MOT+RAMAK(B) is higher than that of LDATM+MOT(A). The results shows that the mean of (B) (M= .191, SD =.062) is significantly higher than the mean of (A) (M= .161, SD =.054), $t(166) = -16.27$, $p < .01$.

Table 6 shows the changes of topical cohesion in major topics, and all topics except T2 had an increase in topical cohesion among the keywords in the topic. The value of within-topic similarity of all topics except T2 increased after applying RAMAK.

The 10 topic categories with the highest increase rate and the 10 topic categories with the lowest increase rate are shown in Table 7. The increase rate of within-topic similarity is a ratio of the within-topic similarity of LDATM+MOT+RAMAK based on the within-topic similarity of LDATM. As shown in Table 7, the increase rate below one is only seven topic categories and the values were close to one.

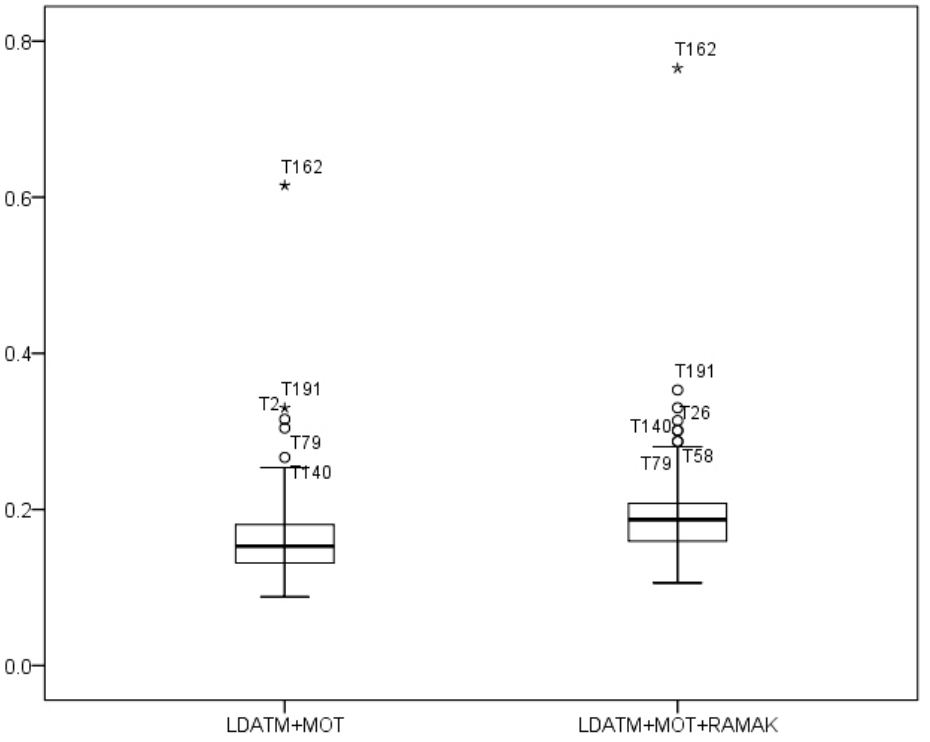


Fig. 8 Difference in within-topic similarity between LDATM+MOT and LDATM+MOT+RAMAK

Table 6. Changes of Topical Cohesion in Major Topics

Topic	Within-Topic Similarity of LDATM+MOT (A)	Within-Topic Similarity of LDATM+MOT+RAMAK (B)	# of Keyword of LDATM+MOT+RAMAK (C)	Increase Rate of Within-Topic Similarity (B/A)
T0	0.200	0.217	16	1.081
T1	0.171	0.200	16	1.167
T2	0.316	0.314	17	0.995
T4	0.330	0.353	12	1.070
T3	0.115	0.139	15	1.205
T5	0.193	0.208	18	1.076
T6	0.122	0.152	13	1.248
T7	0.119	0.152	14	1.283
T9	0.195	0.212	19	1.090

Table 7. Top 10 Topics with the Major Changes in Within-Topic Similarity

Topic	Within-Topic Similarity of LDATM+MOT (A)	Within-Topic Similarity of LDATM+MOT+RAMAK (B)	# of Keyword of LDATM+MOT+RAMAK	Increase Rate of Within-Topic Similarity (B/A)
T179	0.135	0.244	9	1.808
T72	0.098	0.172	10	1.758
T28	0.088	0.145	10	1.650
T148	0.130	0.207	9	1.594
T65	0.141	0.225	10	1.589
T95	0.141	0.221	10	1.565
T14	0.135	0.206	10	1.524
T69	0.132	0.198	13	1.507
T107	0.189	0.280	11	1.486
T86	0.139	0.207	11	1.485
T15	0.206	0.195	21	0.944
T188	0.142	0.135	11	0.957
T143	0.165	0.160	19	0.971
T114	0.112	0.109	18	0.972
T42	0.246	0.240	22	0.977
T79	0.304	0.301	7	0.990
T2	0.316	0.314	17	0.995
T97	0.147	0.148	18	1.002
T189	0.117	0.119	17	1.015
T122	0.131	0.133	19	1.015

Note. The rest of this table is available upon request.

It also implies that a certain topic with fewer top-keywords has a more cohesive structure topically. Spearman's rho was computed between the increase rate of within-topic similarity and the change in the number of keywords. The result of the correlation analysis was statistically significant (r = -.501, p < .01). The result suggested that if the number of keywords in a topic is smaller, the topic tends to have a more cohesive structure among the keywords topically.

4. CONCLUSIONS

This study combined the approaches of natural language processing, text mining, and network analysis to explore the applicability of the concept of synchronization in the result of topic modeling. As a result of applying the combined approach to the domain analysis of construction and building engineering, visibility not only in the relationships of topic-keyword but also

in that of topic-topic was observed as being improved and topical cohesion in a topic was significantly increased overall.

The combined approach of this study could be regarded as being easy to apply with the best of knowledge on the phenomenon of synchronization in a complex network. The approach of re-assigning the multi-assigned keyword after merging similar topics on the result of LDA-based topic modeling is practical and step-wise. Also, the approach suggested in this study could provide more acceptable and interpretable evidence to researchers and experts in a specific domain in order to help them to detect and analyze the overview of the domain based on their knowledge of it, because the approach could reduce complex connections between multi-assigned keywords and topics.

Because the suggested approach of this study is exploratory, there should be more cases of applying the approach in various domains for generalization and advancement of the approach. More sophisticated evaluation, such as performing user evaluation or calculating a ratio of between-topic similarity and within-topic similarity, should also be investigated to develop the advanced methods of the suggested approach.

REFERENCES

Arenas, A., Diaz-Guilera, A., Kurths, J., Moreno, Y., & Zhou, C. (2008). Synchronization in complex networks. *Physics Reports, 469*(3), 93-153.

Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2009, June). On smoothing and inference for topic models. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (pp. 27-34). AUAI Press.

Brown, P. F., Pietra, V. J. D., Mercer, R. L., Pietra, S. A. D., & Lai, J. C. (1992). An estimate of an upper bound for the entropy of English. *Computational Linguistics, 18*(1), 31-40.

Blasius, B., Huppert, A., & Stone, L. (1999). Complex dynamics and phase synchronization in spatially extended ecological systems. *Nature, 399*(6734), 354-359.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Blei, D. M., & McAuliffe, J. D. (2010). Supervised topic models. arXiv preprint arXiv:1003.0783.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Advances in neural information processing systems (pp. 288-296).

Chuang, J., Ramage, D., Manning, C., & Heer, J. (2012). Interpretation and trust: Designing model-driven visualizations for text analysis. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 443-452). ACM.

Elowitz, M. B., & Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature, 403*(6767), 335-338.

Garcia-Ojalvo, J., Elowitz, M. B., & Strogatz, S. H. (2004). Modeling a synthetic multicellular clock: Repressilators coupled by quorum sensing. Proceedings of the National Academy of Sciences of the United States of America, *101*(30), 10955-10960.

Jalili, M. (2013). Enhancing synchronizability of diffusively coupled dynamical networks: A survey. *Neural Networks and Learning Systems, IEEE Transactions on, 24*(7), 1009-1022.

Jha, S. S., & Yadava, R. D. S. (2012). Synchronization based saw sensor using delay line coupled dual oscillator phase dynamics. *Sensors & Transducers* (1726-5479), *141*(6), 71-91.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, *101*(Suppl 1), 5228-5235.

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 363-371). Association for Computational Linguistics.

Kang, B-I., Song, M., Jho, H.S. (2013). A study on opinion mining of newspaper texts based on topic modeling. *Journal of the Korean Library and Information Science Society, 47*(4), 315-334.

Kuramoto, Y., & Nishikawa, I. (1987). Statistical macrodynamics of large dynamical systems. Case of a phase transition in oscillator communities. *Journal of Statistical Physics, 49*(3-4), 569-605.

Lu, Y., & Zhai, C. (2008). Opinion integration through

semi-supervised topic modeling. In Proceedings of the 17th international conference on World Wide Web (pp. 121-130). ACM.

Mirollo, R. E., & Strogatz, S. H. (1990). Synchronization of pulse-coupled biological oscillators. *SIAM Journal on Applied Mathematics, 50*(6), 1645-1662.

Miyano, T., & Tsutsui, T. (2007a). Data synchronization in a network of coupled phase oscillators. *Physical Review Letters, 98*(2), 024102.

Miyano, T., & Tsutsui, T. (2007b). Extracting feature patterns in the health status of elderly people needing nursing care by data synchronization. In Information Technology Applications in Biomedicine, 2007. ITAB 2007. 6th International Special Topic Conference on (pp. 153-156). IEEE.

Miyano, T., & Tsutsui, T. (2008a). Collective synchronization as a method of learning and generalization from sparse data. *Physical Review E, 77*(2), 026112.

Miyano, T., & Tsutsui, T. (2008b). Finding major patterns of aging process by data synchronization. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, *91*(9), 2514-2519.

Miyano, T., & Tsutsui, T. (2009). Link of data synchronization to self-organizing map algorithm. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, *92*(1), 263-269.

Miyano, T., & Tatsumi, K. (2012). Determining anomalous dynamic patterns in price indexes of the London Metal Exchange by data synchronization. *Physica A: Statistical Mechanics and its Applications, 391*(22), 5500-5511.

Miyano, T., & Tsutsui, T. (2007). Data synchronization as a method of data mining. In Proceedings of the 2007 International Symposium on Nonlinear Theory and its Applications NOLTA'07 (pp 224-227). Vancouver: NOLTA.

Niebur, E., Schuster, H. G., Kammen, D. M., & Koch, C. (1991). Oscillator-phase coupling for different two-dimensional network connectivities. *Physical Review A, 44*(10), 6895.

Park, J-H. & Song, M. (2013). A study on the research trends in library & information science in Korea using topic modeling. *Journal of Korean Society for Information Management, 30*(1), 7-32.

Pikovsky, A., Rosenblum, M., & Kurths, J. (Eds.). (2003). *Synchronization: A universal concept in nonlinear sciences* (Vol. 12). London: Cambridge University Press.

Pluchino, A., Latora, V., & Rapisarda, A. (2005). Changing opinions in a changing world: A new perspective in sociophysics. *International Journal of Modern Physics C, 16*(04), 515-531.

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009a). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 (pp. 248-256). Association for Computational Linguistics.

Ramage, D., Rosen, E., Chuang, J., Manning, C. D., & McFarland, D. A. (2009b). Topic modeling for the social sciences. In NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond (Vol. 5).

Ramage, D., Manning, C. D., & Dumais, S. (2011). Partially labeled topic models for interpretable text mining. In Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 457-465). ACM.

Strogatz, S. H., & Mirollo, R. E. (1988). Collective synchronisation in lattices of nonlinear oscillators with randomness. *Journal of Physics A: Mathematical and General, 21*(13), L699.

Strogatz, S. H. (2000). From Kuramoto to Crawford: Exploring the onset of synchronization in populations of coupled oscillators. *Physica D: Nonlinear Phenomena, 143*(1), 1-20.

Strogatz, S. H. (2001). Exploring complex networks. *Nature, 410*(6825), 268-276.

Strogatz, S. (2003). *Sync: The emerging science of spontaneous order*. New York: Hyperion.

Talley, E. M., Newman, D., Mimno, D., Herr II, B. W., Wallach, H. M., Burns, G. A. P. C., Leenders, A. G. M., & McCallum, A. (2011). Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods, 8*(6), 443-444.

Tang, J., Wang, B., Yang, Y., Hu, P., Zhao, Y., Yan, X., & Usadi, A. K. (2012). Patentminer: Topic-driven patent analysis and mining. In Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1366-

1374). ACM.

Tilles, P. F., Cerdeira, H. A., & Ferreira, F. F. (2013). Local attractors, degeneracy and analyticity: Symmetry effects on the locally coupled Kuramoto model. *Chaos, Solitons & Fractals, 49*, 32-46.

Titov, I., & McDonald, R. (2008). Modeling online reviews with multi-grain topic models. In Proceedings of the 17th international conference on World Wide Web (pp. 111-120). ACM.

Wan, M., Li, L., Xiao, J., Yang, Y., Wang, C., & Guo, X. (2010). CAS based clustering algorithm for Web users. *Nonlinear Dynamics, 61*(3), 347-361.

Yu, S.Y. (2013). Applying TDP (Topic Descriptor Profile) with article-level citation flow for analyzing research trend, In proceedings of the 2013 Korean Society for Information Management Conference in Autumn (pp. 39-58). Seoul: Korean Society for Information Management.