Journal of Informetrics 16 (2022) 101255

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Developing a topic-driven method for interdisciplinarity analysis

Hyeyoung Kim, Hyelin Park, Min Song*

Department of Library and Information Science, Yonsei University, Seoul, 03722, Republic of Korea

ARTICLE INFO

Keywords:
Disciplinarity
Interdisciplinary cooperation
Topic diversity
Keyword generation
DMR topic modeling
Deep learning

ABSTRACT

This study explores the topic-based interdisciplinarity in the research domain of literacy. A text corpus of keywords was generated through a deep keyword generation model from abstracts of 346,387 articles published in 296 disciplines from 1917 to 2021. Dirichlet-Multinomial Regression topic modeling, interdisciplinarity indices, and network analysis were employed to analyze the collected corpus. Topic modeling uncovered 15 dominant research topics in the literacy field, as well as their up-and-down trends from 2000 to 2021. For each topic, keywords were then replaced with disciplines, and interdisciplinarity was measured using four indices: variety, balance, disparity, and diversity. Finally, the interdisciplinarity of each topic, connectivity between topics, and topic trends were comprehensively analyzed on the keyword co-occurrence network. Our methodology reaches beyond connectivity limited to a few disciplines and provides insight into the direction of collaboration between disciplines centered on a research domain. Moreover, the study's deep keyword generation model has methodological implications for forming a corpus spanning numerous disciplines as a bottom-up approach.

1. Introduction

Interdisciplinarity research has been commonly used to promote interdisciplinary collaboration by identifying the characteristics of knowledge exchanges and transfer among disciplines. Complex datasets obtained from literature contain information on the current structure and dynamic flow of scientific knowledge. Bibliometric studies have explored interdisciplinarity by employing different units of analysis and applying measurement indices with various perspectives on diversity (Rafols & Meyer, 2007).

The most fundamental condition for exploring interdisciplinarity is ensuring that the dataset in question contains data from two or more disciplines. Leinster and Cobbold (2012) emphasize that the input data for measuring interdisciplinarity should be sufficient and interrelated to properly quantify the disciplines' similarity. The vaster and richer corpus contained by disciplines, the better we can understand the interactions and characteristics of those disciplines. This, in turn, enables the possibility of collaboration to solve common research problems by including more diverse disciplines.

The existing methodology for interdisciplinarity measurement is based on using *journals* as the unit of data analysis (Bu, Li, Gu, & Huang, 2020; Hu & Zhang, 2018; Leydesdorff & Ivanova, 2020). Top-down methods rely primarily on the journal as an analysis unit by utilizing journal-based subject area classification information. The relationships between journals collected from a predefined scheme have been established by different approaches such as journal citation relationships (Leydesdorff & Ivanova, 2020; Leydesdorff, Wagner, & Bornmann, 2018) and betweenness centrality or the cross-citation similarities of journals (Zhang, Janssens, Liang, & Glänzel, 2010; Zhang, Rousseau, & Glänzel, 2016, 2016).

While top-down analysis is conducted by journal unit within a pre-determined discipline classification scheme, the bottom-up method creates topics from keyword units contained in an unspecified research domain. The present study proposes such a bottom-up

* Corresponding author.
E-mail addresses: youngdrewa@yonsei.ac.kr (H. Kim), lin0@yonsei.ac.kr (H. Park), min.song@yonsei.ac.kr (M. Song).

https://doi.org/10.1016/j.joi.2022.101255
Received 20 May 2021; Received in revised form 28 November 2021; Accepted 22 January 2022
Available online 1 February 2022

method to analyze the relationships between disciplines based on keywords. However, in many cases, publications do not contain the author-designed keywords. In addition, this approach requires a more advanced methodology that can reveal the semantic meaning embedded in a document's content from only collecting the keywords that are author-designed or appearing in the source text (Alzaidy et al., 2019). Therefore, for this study, we constructed a keyword generation model via a deep-learning algorithm (Meng et al., 2017) that can generate keywords containing semantic meanings learned from a large-scale document collection. This model enables a corpus to effectively contain keywords summarizing research topics from a large number of disciplines in a bottom-up approach.

To measure topic-based interdisciplinarity, we used four indices: *variety, balance, disparity,* and *diversity*. Indices measuring interdisciplinarity have been developed based on studies in biology, such as Rao (1982) and Stirling (1998), to analyze the diversity of species within specific categories. Rafols and Meyer (2010) proposed a methodological framework that can measure interdisciplinarity in bibliometric studies based on diversity indices such as variety, balance, and disparity; i.e., the three concepts of Rao-Stirling diversity. As such, these three indices have been mainly used to measure interdisciplinary in top-down methods using journals as media. On the other hand, diversity is an index that has been applied to bottom-up studies that measure based on topics formed using keywords as media (Webber et al., 2010).

Combined with a keyword generation method based on deep learning, this study employs Dirichlet-Multinomial Regression (DMR) topic modeling (Mimno & McCallum, 2008) to detect topics embedded in the literature and to analyze topic trends by year using keyword probabilities. Based on these results, we measured interdisciplinarity using the four selected indices. Moreover, each topic's interdisciplinarity and the relationships between topics were comprehensively analyzed through the constructed network based on keyword co-occurrence.

For those interested in the proposed method and for a wide dissemination of the new approach, we made the Python-based code and the experiment results publicly available at https://github.com/HyelinPark/Topic-Based-Interdisciplinarity-Analysis.

The rest of this paper is organized as follows. We first review related work in Section 2. Then, we present the study's methodology in Section 3. We describe the results in Section 4, which are followed by our discussion in Section 5. Section 6 concludes the paper.

2. Related work

2.1. Interdisciplinarity

Interdisciplinarity can be defined as the integration of information, data, techniques, tools, perspectives, concepts, and theories from two or more disciplines, as well as an attempt to identify solutions that solve a problem beyond a single discipline or domain of research practice (National Academies Committee on Facilitating Interdisciplinary Research, 2005).

Interdisciplinary research is necessary to create opportunities for disparate disciplines to engage with each other (Hicks et al., 2010). Input data for measuring interdisciplinarity must be sufficiently abundant to be able to quantify similarities between disciplines (Leinster & Cobbold, 2012). When a corpus is limited to a specific discipline, analyzing the relationships among more vast disciplines is impossible. To resolve this issue, Hu and Zhang (2018) constructed a comprehensive corpus by including 109 disciplines in their Big Data research domain, making it possible to effectively measure interdisciplinarity covering a large number of disciplines.

Methods for measuring interdisciplinarity have been developed based on the concepts presented by Rao (1982) and Stirling (1998, 2007), which calculate numbers, similarities, or balances among categories. Most research has attempted to develop improved indices based on the Rao-Stiring index (Zhang et al., 2010; Rousseau, 2018; Zhang, Rousseau, & Glänzel, 2016). Therefore, in this study, we apply the basic three indices: variety, disparity, and balance. Each index exhibited different disciplinary characteristics; namely, the variety index was able to identify how many disciplines interact in each research topic, while balance and disparity identified how many heterogeneous disciplines were evenly distributed with common topics.

Efforts have been made to develop improved diversity indices by further integrating the concepts of balance and similarity. For instance, Zhang et al. (2010) improved subject classification schemes via clustering based on journal citation similarities. Zhang et al. (2016) found that the Rao-Stirling measure has low discriminatory power. Accordingly, they created a journal cross-citation matrix for seven selected journals, then assigned them to a subject classification scheme, and finally measured interdisciplinarity by applying the diversity concept. Rousseau (2018) also indicated that Rao-Stirling diversity did not meet the "monotonicity of balance" requirement and sought an alternative method to measure that balance. Bu et al. (2020) collected keywords extracted from abstracts with the TextRank algorithm and measured journals' interdisciplinarity using topic diversity, a topic detection method for extracting fine-grained topics.

Topic diversity in this study is measured by calculating ranking-biased overlap (RBO), which is capable of comparing lists that are incomplete or that have only some members in common and revealing how similar disciplines are shared across other topics (Webber et al., 2010). Therefore, this study explores four different interdisciplinary characteristics by applying four indices: variety, similarity, disparity, and diversity.

Based on each perspective of the four indices of variety, balance, disparity, and diversity, we redefined interdisciplinarity. In this paper, interdisciplinarity is defined as: 1) the diversity of disciplines; i.e., a greater number of disciplines or a more balanced ratio between disciplines, and 2) the similarity (not heterogeneity) of keywords shared between disciplines.

2.2. Top-down approaches

Top-down studies have measured the disciplinary diversity of knowledge systems composed of predefined categories (i.e., subject classification schemes). Most studies have measured journals' interdisciplinarity by applying a top-down method based on subject classifications such as the Web of Science (Asubiaro & Badmus, 2020) and the Social Sciences Citation Index (Leydesdorff & Goldstone, 2014). Leydesdorff's studies (Leydesdorff, 2007; Leydesdorff & Goldstone, 2014; Leydesdorff, Wagner, & Bornmann, 2018, 2019; Leydesdorff & Ivanova, 2020) measured interdisciplinarity as a characteristic of journals. Primarily, the above studies constructed a journal-journal citation network based on journals' citation similarities and measured interdisciplinarity by applying network coherence indicators (i.e., betweenness centrality) and diversity indices (i.e., the Rao-Stirling index). Other studies have applied top-down methods that measure interdisciplinarity using journals as an analysis unit by acquiring discipline information from journal-based subject categories (Hu & Zhang, 2018; Zhang et al., 2010, 2016).

Since these studies use existing, already established systems, the boundaries between research subject areas are clear. While this has the advantage of making it relatively easy to construct corpus of different yet comparable disciplines, it is impossible to include in the corpus a large number of disciplines that do not fall within the category of journals set as the scope of analysis.

2.3. Bottom-up approaches

To overcome the limitation of existing interdisciplinarity measurement indices—which are mainly based on the co-existence of authors, institutions, or citations—a topic-based interdisciplinarity measurement method has been proposed by Xu et al. (2016). However, some limitations still remain. First, Xu et al. (2016) study analyzed keywords acquired within a specific discipline, Information Science & Library Science (LIS). Accordingly, relationships between more diverse disciplines, centered on the highly interdisciplinary research domain outside the LIS, were not considered. Second, they employed the Thomson Data Analyzer (TDA) as their keyword collection method, which tokenizes the original word but does not make it possible to confirm whether a large amount of keywords with more semantic content have been sufficiently obtained from the literature. Third, in terms of topic formation, they considered high-frequency terms in LIS as topics. Then, interdisciplinarity was measured by discovering the intersections of internal topics and external disciplines of the LIS. However, this topic formation method is limited in detecting topics hidden in the corpus compared to the topic modeling method that considers the appearance probability based on the simultaneous use pattern of keywords in the literature. Fourth, their method for measuring interdisciplinarity was based on keyword frequency. The interdisciplinarity measurement method that takes into account various aspects of the existing interdisciplinarity indices (e.g., Rao-Stirling's variety, balance, topic diversity, information entropy) was not applied; thus, the similarity between keywords within the topic was not analyzed in greater depth.

Bottom-up studies have measured the interdisciplinarity of large numbers of disciplines by applying text-mining techniques using keywords. For instance, Bu et al. (2020) composed a research topic with keywords extracted via a graph-based ranking algorithm by PageRank called TextRank (Mihalcea & Tarau, 2004). As a technique for organizing topics through a bottom-up method with keywords, many studies have applied topic modeling approaches. Suominen and Toivanen (2016) have argued that automated classification schemes are highly dependent on input data but have more potential value in identifying emerging research topics than historical classification. Thus, they proposed a machine-learning classification scheme as an LDA method that can read latent patterns from data abstracts. In this study, we applied the Dirichlet-Multinominal Regression (DMR) topic model, which can input the publication year as metadata to identify topic extraction and topic trends rising or falling.

As relationships between topics have been effectively detected in the co-work network, the issue of whether it is necessary to apply citation relationships has been raised (Bu et al., 2020). Meanwhile, citation relationships between journals have been widely applied (Leydesdorff & Ivanova, 2020). As a way to establish relationships between heterogeneous systems, Zhou et al. (2012) built a framework that enables factor analysis to build a multidimensional scaling (MDS) map based on categories' weighted similarity. Rafols and Meyer (2010) consider the bottom-up method to be different from the top-down method not because it does not use predefined categories, but rather because knowledge integration is newly formed and analyzed as network coherence indicators.

Therefore, we have included a large number of disciplines in the literacy research domain into our corpus. Our main unit of analysis is keywords generated from abstracts of numerous literatures through a deep keyword model. This model allows us to obtain deep semantic keywords regarded as the core thematic information of a longer text with a state-of-the-art technique (Meng et al., 2017). Based on the topics detected through the topic modeling technique, interdisciplinarity was comprehensively analyzed through the network constructed via keyword co-occurrence.

3. Methodology

In this section, the study's proposed approaches are described and illustrated in Fig. 1. These are comprised of four steps: keyword generation, topic modeling, interdisciplinarity measurement, and topic-based interdisciplinarity analysis based on the keyword co-occurrence network. Detailed processes for each step are as follows.

3.1. Deep keyword generation model

Using the Scopus API, abstracts and several metadata of documents containing the word 'literacy' in titles were collected without any subject area limitation. The search scope was limited to article and conference papers as document type and English as the language. We collected abstracts, titles, and author-designed keywords to obtain input data for topic modeling. Additionally, year of publication was collected to analyze the yearly distribution of topics. Then, in order to replace keywords with disciplines, subject area metadata were acquired based on the All Science Journal Classification Codes (ASJC), which Scopus uses to classify journals (see Appendix A). This classification features an in-depth scheme of disciplines including domain, field, and subfield. To analyze

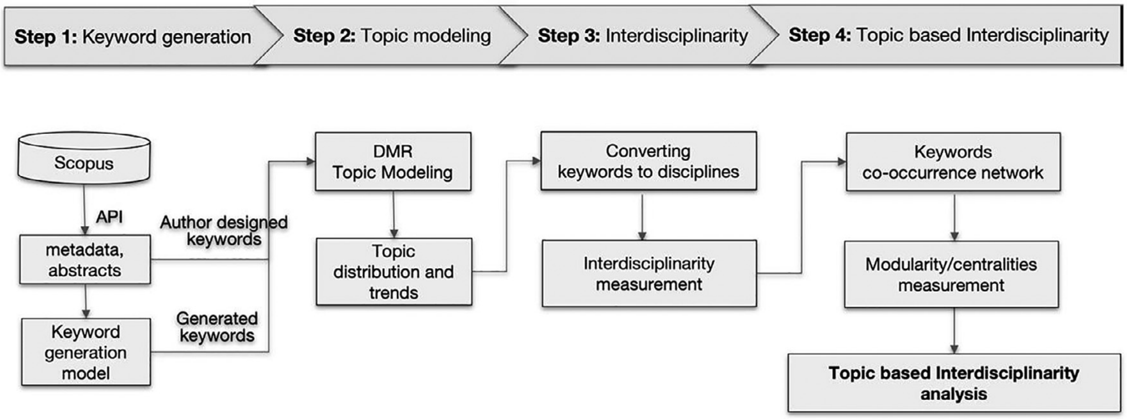

Step 1: Keyword generation
Step 2: Topic modeling
Step 3: Interdisciplinarity
Step 4: Topic based Interdisciplinarity
Scopus
API
metadata, abstracts
Keyword generation model
Author designed keywords
Generated keywords
DMR Topic Modeling
Topic distribution and trends
Converting keywords to disciplines
Interdisciplinarity measurement
Keywords co-occurrence network
Modularity/centralities measurement
Topic based Interdisciplinarity analysis

Fig. 1. Overview of our research method.

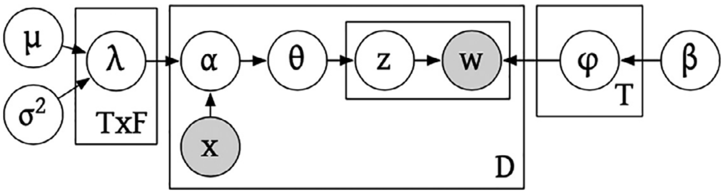


Fig. 2. Graphical model representation of DMR. Source: Mimno and McCallum (2008).

more specific and larger numbers of subject areas, we selected ‘subfield’ for discipline information. Some irregular subfields were integrated; for example, “Medicine (all)” and “Medicine (miscellaneous)” were merged into “Medicine”.

To collect keywords from abstracts and titles using a bottom-up approach, we adopted a deep keyword generation model. This model represents the semantic meaning of a given text by using an RNN Encoder-Decoder model (also referred to as sequence-to-sequence learning) with a copying mechanism (Cho et al., 2014; Gu et al., 2016). Basically, it follows the implementation of the RNN Encoder-Decoder model. The encoder compresses keywords along with semantic information into a hidden representation, while the decoder creates a variable-length sequence by decompressing it with a dense vector. This encoder-decoder network is trained until the conditional probability of the target sequence is optimized. Moreover, by incorporating a copying mechanism, RNN can predict out-of-vocabulary words by selecting suitable words from texts and allowing the copying mechanism to weigh the importance of each word with a measure of positional attention. This is important because significant phrases can also be identified by positional and syntactic information in their context. Since the CopyRNN is executed by predicting out-of-vocabulary words, it is possible to generate topically relevant keywords that are absent from the text based on deep semantic meaning (Meng et al., 2017).

Automatic keyword extraction algorithms have been widely used in bottom-up approaches to acquire keywords from texts. This method produces a list of candidate keywords from the text and divides them into multiple text chunks using n-grams or noun phrases. Then, keywords are ranked with machine-learning features such as TF-IDF and PageRank, and the most meaningful keywords are selected. However, this method is limited in capturing semantic contents from documents because it can only acquire keywords appearing in the text and evaluates importance based on word occurrence and co-occurrence. Therefore, in the present study, a state-of-the-art deep keyword generation model was applied to expand the input dataset for topic modeling as a rational approach. To train a keyword generation model based on deep learning, the KP20K dataset (Meng et al., 2017) was used. This dataset consists of 530,809 articles for training and 20,000 articles for testing and validation. Each data unit includes a paper’s title, abstract, and keywords.

3.2. Dirichlet-Multinomial regression (DMR) topic modeling

This study constructed topics utilizing a bottom-up method with keywords as the input, aiming to analyze a timeline of rising and falling topic trends after the 2000s. Topic modeling is an algorithm that can be used to group words that are likely to appear in the same context in a vast unstructured literature group. Dirichlet-Multinomial Regression (DMR) topic modeling (Mimno & McCallum, 2008) is a technique used for deriving topic results by setting articles’ metadata features—such as authors, publishers, references, and date information—as third parameters on the basis of literature and topic distribution.

In the model in Fig. 2, the initial parameter value α is determined by λ, which is one-hot encoded into an F-dimensional vector by multiplying the number of topics T and the number of metadata F (TxF). The α depends on metadata x and determines the topic distribution of the document θ. Subsequently, topic z is vectorized. The observed words w is calculated according to word distribution by topics φ, and then adjusted by the hyper parameter value β. We selected published year as metadata x and used it for the document specific Dirichlet parameter.

This study applied DMR topic modeling analysis to the corpus, which consists of keywords obtained from 346,387 papers published in 296 disciplines from 1917 to 2021. During the topic modeling process, we found duplicate keywords in all topics generated by topic modeling. Therefore, we observed the discrimination of keywords by topic while adjusting the upper frequency threshold in multiple units of five. As a result, when deleting 40 points, it was possible to obtain differentiated keywords between topics while having little impact on the value of perplexity. The top 40 high-frequency words removed are listed in Appendix B.

A strong advantage of DMR topic modeling is that it can identify distribution trends using articles' metadata features. We set publication year as a parameter to understand the annual trends of topics from 2000 to early 2021. Understanding the rising and falling patterns of topics from the 2000s to the present helps to predict topics' future research trends. Therefore, a large amount of data published between 1917 and 2021 was used in this study to model the DRM topic. However, only the period from 2000 to 2021 was visualized for yearly trend analysis. Data for 2021 was collected around February 2021, and although remarkably small, it has been included to capture trends up to this point in time as much as possible. It occupies a very small percentage of the total data and does not affect the overall topic distribution. However, in 2021, as a large number of related publications were published due to the impact of the COVID-19 pandemic, it may affect the topic distribution in the overall literacy domain. Therefore, it is necessary to add data for a more complete analysis of the distribution of topics in 2021.

3.3. Interdisciplinarity measurement

In this step, four interdisciplinarity indices were calculated, and the results were applied to the topic-based network constructed in step 3.3. The three indices of variety, balance, and disparity were calculated by applying the formula developed from the research of Rao (1982) and Stirling (1998; 2007). The fourth index, the topic diversity index (Webber et al., 2010), measured topic similarities using the rank-biased overlap (RBO) of each topic's keywords.

Variety

When constructing the topics, the discipline information mapping with 15 keywords was collected and the number of their unique non-redundant disciplines was used as the variety value. For example, if there was a topic consisting of two keywords—'literacy' and 'information'—and the keywords were mapped with three documents from corresponding disciplines—'medical', 'medical', and 'computer science', respectively—then the topic's variety value would be 2.

Balance

After collecting the discipline information replaced with 15 keywords when constructing the topics, the number of disciplines and the balance of frequencies between disciplines were calculated using the Gini coefficient. This coefficient is a measure of inequality defined as the mean of absolute differences between all pairs of individuals for some measure Stuart et al., 1994). The closer the value is to 0, the more equal it is; the closer it is to 1, the greater the gap in the ratio of appearances between academic fields. Where x is an observed value, n is the number of values observed, and i is the rank of values in ascending order, the Gini coefficient is shown in Eq. (1) and ((2).

$$G = \frac{2}{n^2 \bar{x}} \sum_{i=1}^{n} i(x_i - \bar{x}) \tag{1}$$

$$G = \frac{\sum_{i=1}^{n} (2i - n - 1)x_i}{n \sum_{i=1}^{n} x_i} \tag{2}$$

Disparity

After obtaining the entire set of keywords that appeared in each discipline, we obtained keyword-based similarities by disciplines through a cosine similarity calculation. The average similarity between all disciplines where topic keywords appeared was subtracted from 1 and employed as the disparity value. When we refer to the list of keywords by disciplines as vectors A and B, disparity is expressed as (3).

$$Disparity = 1 \;-\; similarity = 1 - \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} \left(A_i\right)^2} \times \sqrt{\sum_{i=1}^{n} \left(B_i\right)^2}} \tag{3}$$

Topic diversity

Each topic's diversity was measured through rank-biased overlap (RBO). RBO is a similarity measure appropriate for indefinite rankings that measures the score of the rank-biased overlap over the topics. It is top-weighted, handles non-conjointness in the rankings, and is free from any particular prefix length (Webber et al., 2010). RBO compares two ranked lists and returns a value between zero and one to quantify their similarity; a value of 0 indicates that the lists are completely different. The sequence of the lists of disciplines that make up each topic depends on the ranking of the probability values of the keywords contributed to topic formation. Therefore, topic diversity values are closely connected to the characteristics of the keywords that make up a particular discipline. Following the equation in Eq. (5), it can be determined that topics with a low topic diversity value (high RBO value) are comprised of keywords that are frequently shared with many other disciplines or general concept keywords. Thus, this means that such topics have a relatively high degree of interdisciplinarity. On the other hand, topics featuring a high topic diversity value (low RBO value) can be presumed to consist of more specific technical terms or concept keywords that are primarily studied in a few particular disciplines, meaning that such topics have a relatively low degree of interdisciplinarity.

We conducted calculations on 15 topics, considering the list of academic fields to which each topic's keywords belong as a ranked list. Setting $p = 0.9$ for user persistence in Eq. (4), and given two ranked lists S and T, d denotes top d documents in S and T. The overlap between S and T is defined at depth d as the size of the intersection between these lists at depth d and divided by the depth; topic diversity is expressed as the inverted mean value of RBO.

$$RBO(S, T, p) = (1 - p) \sum_{d=1}^{\infty} p^{d-1} \cdot A_d \tag{4}$$

$$Topic\ Diversity = 1 - mean(RBO) \tag{5}$$

All four indexes employed in this study provide different information for identifying multidisciplinary topics. High interdisciplinary topics can be discovered and interpreted by comprehensively considering all aspects of each of these indexes.

3.4. Network construction and analysis

In this study, we calculated keyword co-occurrence based on the topic modeling results. The degree of connection between keywords was quantitatively analyzed by converting the two-mode network of topic-keyword into the one-mode network of topic-topic. In the network, the nodes represent 15 topics while the edges represent co-occurrence relationships among nodes.

These methods center around network analysis (Friedkin, 1991; Wolfe, 1997) by analyzing centralities to identify important nodes in the network. Betweenness centrality measures the degree of nodes that act as mediators in the entire network. The high centrality of a specific node (i.e., topic) means that topic has a high influence so that other topics appear together. In this study, 15 topics are represented by nodes and betweenness centrality is applied as node sizes. UCINET (Borgatti et al., 2002) was used for a topics-keywords matrix construction and centrality calculations. Meanwhile, Excel and R programming were used for data processing, and Gephi (Bastian et al., 2009) was employed as the study's graph-based visualization software. For graph layout, we utilized the Fruchterman-Reingold layout, which belongs to the family of force-directed graph layout algorithms. For the detection of clusters within the network, we used Gephi's modularity tool, which relays the algorithm from Blondel et al. (2008), also known as the Louvain method.

4. Results and analysis

We conducted an empirical study on three different tasks to analyze interdisciplinarity focusing on research topics.

4.1. Data collection

In order to form a corpus that encompasses various disciplines, we chose *literacy* as our target research domain. The literacy domain is comprised of diverse topics with terms such as "digital literacy, information literacy, data literacy, media literacy, early literacy, health literacy, and physical literacy" (e.g., Billington, 2016; Boechler, Dragon, & Wasniewski, 2014; Bröder et al., 2017; Owusu-Ansah, 2005). It is a vastly multidisciplinary research field that has been actively discussed while sharing perspectives across multi-disciplines such as library and information science, education, linguistics, computer science, medicine, etc.

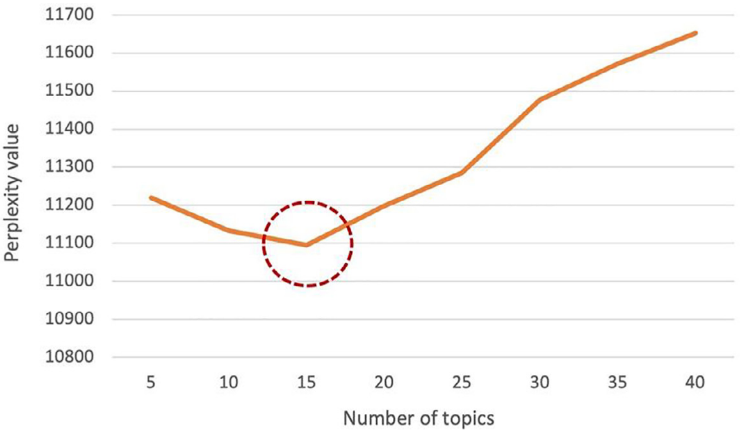
Searching for 'literacy', author-designed keywords and abstracts were collected from 346,387 publications through the Scopus API. Most articles were missing author keywords, as Table 2 shows. By constructing a keyword generation model, we obtained not only author keywords but also more semantic and abundant keywords (Table 3). Finally, we formed a corpus by collecting keywords from abstracts in 346,387 articles published in 296 disciplines from 1917 to 2021.

4.2. Topic distribution and trends

Topic distribution

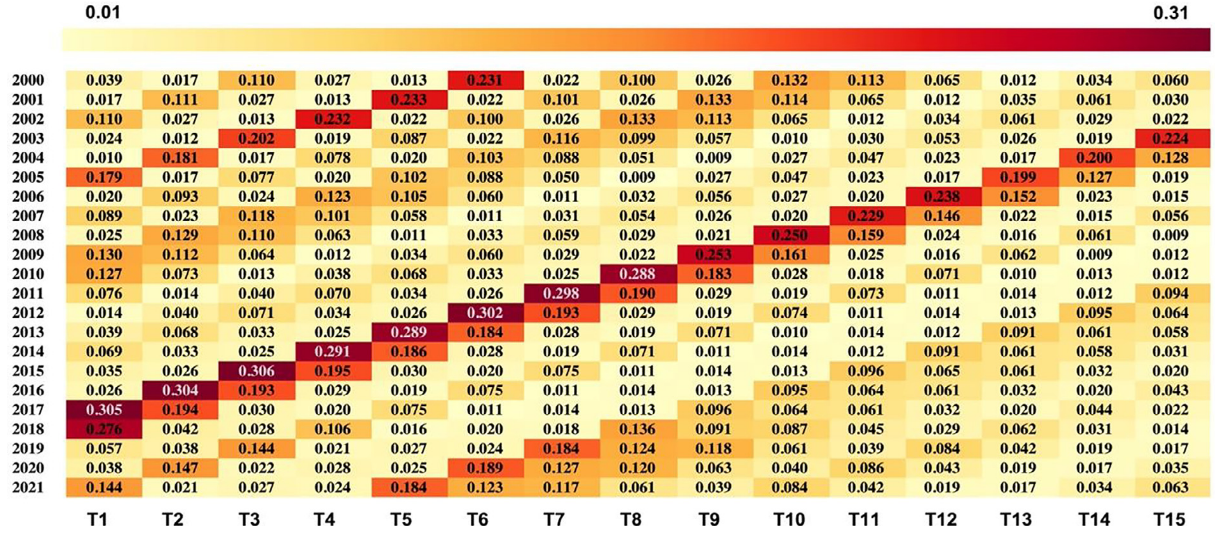
Through DMR topic modeling, we discovered the major topics forming the corpus. Through about 2000 iteration experiments, the number of topics with the lowest perplexity value among at least five to 40 topics was found. As Fig. 3 shows, the number of 15 topics was derived to most appropriately reveal the corpus' topics. The 15 identified topics and the arrangement of keywords with high probability values for each topic are summarized in Table 4. Each topic shared keywords with overlapping meanings, but also had unique keywords that were not found in other topics. These results allowed us to identify the knowledge structure that forms the literacy research domain through 15 topics.

Prior to labeling, our researchers reviewed the classification of literacy research domains in the literature. In previous studies, the literacy domain has been divided into "digital literacy, information literacy, data literacy, media literacy, early literacy, health literacy, and physical literacy" (e.g., Boechler et al., 2014; Bröder et al., 2017; Billington, 2016; Owusu-Ansah, 2005). Thereafter, labels were attached to each topic. The labeling guidelines for each topic were established as follows. First, among the keywords listed in the order of probability values in Table 4, the keyword with the highest probability value was given priority. Second, when the



Number of topics	Perplexity value
5	11220.06
10	11133.51
15	11095.16
20	11198.07
25	11285.56

Fig. 3. Perplexity comparison by number of topics.



	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
2000	0.039	0.017	0.110	0.027	0.013	0.231	0.022	0.100	0.026	0.132	0.113	0.065	0.012	0.034	0.060
2001	0.017	0.111	0.027	0.013	0.233	0.022	0.101	0.026	0.133	0.114	0.065	0.012	0.035	0.061	0.030
2002	0.110	0.027	0.013	0.232	0.022	0.100	0.026	0.133	0.113	0.065	0.012	0.034	0.061	0.029	0.022
2003	0.024	0.012	0.202	0.019	0.087	0.022	0.116	0.099	0.057	0.010	0.030	0.053	0.026	0.019	0.224
2004	0.010	0.181	0.017	0.078	0.020	0.103	0.088	0.051	0.009	0.027	0.047	0.023	0.017	0.200	0.128
2005	0.179	0.017	0.077	0.020	0.102	0.088	0.050	0.009	0.027	0.047	0.023	0.017	0.199	0.127	0.019
2006	0.020	0.093	0.024	0.123	0.105	0.060	0.011	0.032	0.056	0.027	0.020	0.238	0.152	0.023	0.015
2007	0.089	0.023	0.118	0.101	0.058	0.011	0.031	0.054	0.026	0.020	0.229	0.146	0.022	0.015	0.056
2008	0.025	0.129	0.110	0.063	0.011	0.033	0.059	0.029	0.021	0.250	0.159	0.024	0.016	0.061	0.009
2009	0.130	0.112	0.064	0.012	0.034	0.060	0.029	0.022	0.253	0.161	0.025	0.016	0.062	0.009	0.012
2010	0.127	0.073	0.013	0.038	0.068	0.033	0.025	0.288	0.183	0.028	0.018	0.071	0.010	0.013	0.012
2011	0.076	0.014	0.040	0.070	0.034	0.026	0.298	0.190	0.029	0.019	0.073	0.011	0.014	0.012	0.094
2012	0.014	0.040	0.071	0.034	0.026	0.302	0.193	0.029	0.019	0.074	0.011	0.014	0.013	0.095	0.064
2013	0.039	0.068	0.033	0.025	0.289	0.184	0.028	0.019	0.071	0.010	0.014	0.012	0.091	0.061	0.058
2014	0.069	0.033	0.025	0.291	0.186	0.028	0.019	0.071	0.011	0.014	0.012	0.091	0.061	0.058	0.031
2015	0.035	0.026	0.306	0.195	0.030	0.020	0.075	0.011	0.014	0.013	0.096	0.065	0.061	0.032	0.020
2016	0.026	0.304	0.193	0.029	0.019	0.075	0.011	0.014	0.013	0.095	0.064	0.061	0.032	0.020	0.043
2017	0.305	0.194	0.030	0.020	0.075	0.011	0.014	0.013	0.096	0.064	0.061	0.032	0.020	0.044	0.022
2018	0.276	0.042	0.028	0.106	0.016	0.020	0.018	0.136	0.091	0.087	0.045	0.029	0.062	0.031	0.014
2019	0.057	0.038	0.144	0.021	0.027	0.024	0.184	0.124	0.118	0.061	0.039	0.084	0.042	0.019	0.017
2020	0.038	0.147	0.022	0.028	0.025	0.189	0.127	0.120	0.063	0.040	0.086	0.043	0.019	0.017	0.035
2021	0.144	0.021	0.027	0.024	0.184	0.123	0.117	0.061	0.039	0.084	0.042	0.019	0.017	0.034	0.063

Fig. 4. Heatmap of topic distribution from 2000 to 2021.

top-ranked keywords overlapped, the labeling was attached to comprehensively reveal the meaning of the topic taking into account the correlation with the topic's other keywords.

Topic trends from 2000 to 2021

The annual trends of 15 topics detected through DMR topic modeling were analyzed. The purpose of the analysis was not to illuminate the historical flow of 15 topics, but rather to predict the flow of future topics by identifying more recent rising and falling patterns. Therefore, historical data before 2000 were excluded from the visualization analysis. The annual heatmap for topic distribution from 2000 to early 2021 is shown in Fig. 4.

From T15 to T1, topics show a tendency to have been studied with recently focused keywords. Game & e-learning (T3), early literacy (T2), and social network education (T1) received attention in that order from 2015 to 2017. These topics are composed of keywords such as 'online learning', 'video games', 'early childhood', 'academic achievement', and 'social network'. This trend can be interpreted as a growing academic interest in the learning achievements of children and adolescents through online game learning and social networking environments. Topics starting to get attention towards 2021 are information literacy (T5), early language literacy (T6), and digital & collaborative learning (T7). These are topics with keywords such as 'ICT', 'technology', 'media literacy', 'language learning', and 'collaborative learning'. This pattern may be interpreted as an increase in the need for early childhood literacy learning with the advent of the digital technology environment.

In this study, only data up to February 2021 have been included. If data after the COVID-19 pandemic are added, the distribution ratio in 2021 is expected to change in public health (T4) and health information (T8). According to our data, the pattern of attention sequentially from topic 1 to topic 15 has been repeated. Topic 1 to Topic 7 can be seen as revealing a growing need for learning in the digital technology and social networking environment. Topic 8 to Topic 15 emphasize specific learning topics (e.g., environment,

Table 1
Meaning of interdisciplinarity measurement within a topic.

Indices	Measures	Results	Meaning of results	Interdisciplinarity
Variety	Number of disciplines within a topic	High	A greater number of unique, non-overlapping disciplines within a topic	High
Balance	The overall proportion of each discipline within a topic	High	The higher inequality in the ratio of each discipline within a topic	Low
Disparity	1 minus the similarity of keyword appearance between disciplines based on the entire set of keywords of disciplines within a topic	High	The higher heterogeneity of overall keywords in disciplines within a topic	Low
Diversity	Heterogeneity calculated by comparing lists between disciplines ranked with keyword probability used in modeling a topic	High	The higher heterogeneity of keywords with high ranking of topic modeling probability in disciplines within a topic	Low

Table 2
Descriptive statistics of author-designed keywords and abstracts.

		Abstracts	
		Contained	Missing
Author-designed keywords	**Contained**	198,263	1912
	Missing	102,265	43,947
Total		**346,387**	

Table 3
A comparison of author-designed keywords and generated keywords from abstracts.

	Author-designed keywords	Generated keywords
Doc1	HIV, Indonesia, Islam, kinships of shame, populist morality, sexual and reproductive health, universal health coverage	universal health care, HIV, health care, universal health coverage, health coverage
Doc2	cellphone, mental illness, psychiatry, smartphone, technology, wearables	smartphone, psychiatry, consumer technology, mental health
Doc3	advertising disclosures, advertising literacy, brand effects, influencer effects, influencer marketing, influencer-generated disclosure, vlog advertising	disclosure, advertising, interference, sponsored content

financial, democracy) or actors (e.g., schools, teachers, and libraries). Therefore, after the need for learning increases with the advent of the digital technology environment, there is a tendency for studies to explore the roles and service directions of the actors who implement them.

4.3. Interdisciplinarity measurement

To measure interdisciplinarity, the keywords distributed across 15 topics were replaced with disciplines based on the results of 4.2. One keyword was found in several disciplines while being included in several topics. In the process of turning keywords into disciplines, the internal structure of the 15 topics changed. Therefore, the keywords constituting the 15 topics were reorganized into various disciplines. Thus, interdisciplinarity was measured on a knowledge structure composed of 15 topics containing a variety of disciplines.

The four interdisciplinary metrics suggested in the methodology section of 3.4 were employed. Table 5 shows the values measured by applying the calculation formula of each index. As summarized in Table 1 , each of the four indices measures different aspects of interdisciplinarity with different metrics. Each topic has a different interpretation of interdisciplinarity based on the conflicting results for each index.

First, the ***variety index*** measures the number of disciplines included in each topic. A high variety index means that there are many unique, non-overlapping disciplines within the topic. Therefore, the higher the variety index, the higher the interdisciplinarity. Topics with a large number of disciplines were related to health, such as Public health (T4) and Health information (T8). Game & e-learning literacy (T3) also featured a large number of disciplines. On the other hand, Methods & materials (T14) included a small number of disciplines, resulting in low interdisciplinarity.

Second, the ***balance index*** measures how evenly the disciplines within each topic are distributed. The higher the balance index, the higher the proportion of specific disciplines is unequal. With a high balance index, Methods & materials (T14) is a topic with low interdisciplinarity biased towards specific disciplines. The topics with a low balance index are Game & e-learning literacy (T3), Public

Table 4
15 topics and keywords from DMR topic modeling.

Topic	Labeling	15 Keywords
T1	Social network & education	social networks, science education, social network, sustainability, privacy, ICT, digital literacy, scientific literacy, online learning, climate change, motivation, information technology, social networking, environmental education, trust
T2	Early literacy	academic achievement, parenting, kindergarten, early childhood, early childhood education, motivation, preschool, executive function, school readiness, intervention, elementary school, parents, middle school, early literacy, leadership
T3	Game & e-learning literacy	online learning, collaborative learning, games, game theory, virtual reality, design, video games, multimedia, game design, blended learning, serious games, mobile learning, digital literacy, e-learning, language learning
T4	Public health literacy	diabetes, adherence, depression, type 2, older adults, diabetes mellitus, hypertension, healthcare, medication adherence, HIV, health education, quality of life, nutrition, obesity, public health
T5	Information literacy	theory, e-commerce, cybernetics, information technology, design, computer-mediated communication, ICT, Facebook, IT, e-government, China, information, languages, history, identity
T6	Early language literacy	phonological awareness, dyslexia, kindergarten, vocabulary, music, intervention, speech, comprehension, early literacy, language learning, spelling, emergent literacy, working memory, language impairment, discourse
T7	Digital & collaborative learning	ICT, science education, early childhood education, collaborative learning, pedagogy, teachers, educational technology, digital literacy, interactive learning environments, online learning, media literacy, early childhood, case study, adult learning, identity
T8	Health information literacy	cancer, patient education, healthcare, readability, health information, breast cancer, informed consent, asthma, adherence, depression, nursing, decision making, medical informatics, health education, public health
T9	Democracy literacy	identity, discourse analysis, politics, women, religion, globalization, discourse, history, youth, leadership, language policy, democracy, rhetoric, migration, ethnography
T10	Environment & financial literacy	sustainability, climate change, financial literacy, science education, environmental education, scientific literacy, engineering education, citizen science, sustainable development, climate, youth, biodiversity, financial education, environmental literacy, high school
T11	Language obstacles literacy	dyslexia, working memory, phonological awareness, executive function, speech, developmental dyslexia, brain, autism, auditory processing, attention, phonology, cognition, deaf, intervention, speech perception
T12	Schools & teachers	identity, pedagogy, science education, discourse analysis, social justice, middle school, discourse, teachers, secondary school, diversity, teacher, teacher preparation, multimodal, equity, high school
T13	Library & computer literacy	academic libraries, libraries, financial literacy, library, information technology, librarians, nursing, information, computer literacy, library instruction, computer-mediated communication, information seeking, public libraries, library services, students
T14	Methods & materials	comprehension, methods and materials, digital, media literacies, instructional strategies, teaching/learning strategies, theoretical perspectives, engagement, 4-adolescence, 2-childhood, 3-early adolescence, strategies, motivation, instructional strategies, methods and materials, content literacy
T15	Health literacy for women	cancer, women, breast cancer, cervical cancer, HIV, pregnancy, obesity, depression, public health, health education, cancer screening, health promotion, screening, fertility, colorectal cancer

Table 5
Interdisciplinarity indices by topic.

Topic	Label	Variety	Balance	Disparity	Diversity
T1	Social network & education	181	0.335	**0.236**	0.614
T2	Early literacy	178	0.346	0.241	**0.620**
T3	Game & e-learning literacy	**189**	**0.328**	0.263	0.617
T4	Public health literacy	**195**	**0.328**	**0.304**	**0.620**
T5	Information literacy	181	0.351	0.273	**0.609**
T6	Early language literacy	171	0.358	**0.232**	**0.600**
T7	Digital & collaborative learning	178	0.351	0.246	**0.603**
T8	Health information literacy	**190**	**0.326**	0.263	0.613
T9	Democracy literacy	185	0.333	0.246	0.614
T10	Environment & financial literacy	184	0.338	0.254	0.617
T11	Language obstacles literacy	180	0.341	**0.238**	0.617
T12	Schools & teachers	187	0.345	**0.290**	0.610
T13	Library & computer literacy	187	**0.326**	0.242	0.614
T14	Methods & materials	**137**	**0.473**	**0.297**	**0.605**
T15	Health literacy for women	176	0.364	0.268	**0.627**

Table 6
Topic-based interdisciplinarity values by six clusters.

Cluster	Topic	Label	Variety	Balance	Disparity	Diversity	DegreeCentrality	BetweennessCentrality
#1	T4	Public health literacy	**195**	**0.328**	**0.304**	**0.620**	10.000	3.400
	T8	Health information literacy	**190**	**0.326**	0.263	0.613	11.000	2.833
	T15	Health literacy for women	176	0.364	0.268	**0.627**	10.000	4.400
#2	T6	Early language literacy	171	0.358	**0.232**	**0.600**	11.000	8.329
	T2	Early literacy	178	0.346	0.241	**0.620**	11.000	11.033
	T11	Language obstacles literacy	180	0.341	**0.238**	0.617	6.000	0.000
#3	T14	Methods & materials	**137**	**0.473**	**0.297**	**0.605**	4.000	9.178
#4	T1	Social network & education	181	0.335	**0.236**	0.614	17.000	13.571
	T3	Game & e-learning literacy	**189**	**0.328**	0.263	0.617	7.000	1.282
	T10	Environment & financial literacy	184	0.338	0.254	0.617	10.000	2.889
#5	T5	Information literacy	181	0.351	0.273	**0.609**	10.000	5.250
	T13	Library & computer literacy	187	**0.326**	0.242	0.614	6.000	7.317
#6	T7	Digital & collaborative learning	178	0.351	0.246	**0.603**	15.000	2.093
	T9	Democracy literacy	185	0.333	0.246	0.614	9.000	13.441
	T12	Schools & teachers	187	0.345	**0.290**	0.610	13.000	3.986

health (T4), Health information (T8), and Library & computer literacy (T13), which exhibit high interdisciplinarity. These results are similar to the values of the variety index.

Third, the ***disparity index*** is measured by subtracting from 1 the similarity of keyword appearances between disciplines based on the entire set of keywords included in each topic. A high disparity index indicates low interdisciplinarity because the keywords of disciplines within a topic show high heterogeneity. With low disparity values, Social network & education (T1), Early language literacy (T6), and Language obstacle literacy (T11) correspond to high interdisciplinarity. These topics exhibit a new high interdisciplinarity that was not previously discovered by the variety or balance index. On the other hand, the topics with low interdisciplinarity are Public health (T4), Schools & teachers (T12), and Methods & materials (T14). In particular, Public health (T4) exhibits results that conflict with the variety and balance indices. Thus, T4 can be interpreted as a topic with a large number of disciplines and an even ratio between disciplines, but where the overall discipline keywords are different and specialized.

Fourth, the ***topic diversity index*** measures heterogeneity between disciplines by comparing lists ranked by keyword probability used in topic modeling. The higher the diversity index, the higher the heterogeneity of the top-ranked keywords, which corresponds to a low interdisciplinarity. Due to their low diversity values, Information literacy (T5), Digital & collaborative learning (T7), and Early language literacy (T6) correspond to high interdisciplinarity. Method & materials (T14) exhibits a particularly high interdisciplinarity, in contrast to the results of the other three indices. As a result of comprehensive analysis, T14 was found to have a small number of disciplines and an unbalanced ratio of disciplines, where the overall discipline keywords are different. However, on the other hand, high-ranking keywords show high similarity between disciplines. In contrast, Public health (T4) exhibits the same low interdisciplinarity as in the disparity index result. Therefore, in T4, not only overall keywords but also keywords with high-ranking values are very heterogeneous and specialized, meaning that they are not shared across disciplines.

4.4. Topic-based interdisciplinarity analysis

In 4.2, we revealed the major topics of the literacy domain corpus and the annual trends of topics through DMR topic modeling analysis. Then, in 4.3, the interdisciplinarity of each detected topic was measured using four indices. In this section, we present our proposed network, which can show the relationship between topics, and comprehensively analyze topic-based interdisciplinarity through this network.

Relationships of topics based on keyword co-occurrence

To analyze the relationships between topics, a network was constructed by calculating the keyword co-occurrence. Each node features 15 topics; when keywords belonging to each topic appear simultaneously, the topics are linked to each other. The higher the keyword co-occurrence, the greater the link weight. Then, node centrality was analyzed, and clusters were detected using the modularity tool (see the results in Appendix C). Modularity is a scale value that measures the density of relations inside clusters to relations outside clusters. This is a method of suggesting an optimized numerical value so that inside clusters have a large number of connections and outside clusters have a small number of connections (García-Vallejo et al., 2019). In our network, the modularity value is 0.415, and three different size clusters were found.

Fig. 5 is a network visualization of analysis results via Gephi software. The nodes are the 15 topics, and the edge between nodes is the keyword co-occurrence. Moreover, the node size represents variety (number of disciplines), and the node color is the three clusters detected by the modularity tool.

Comprehensive analysis by six clusters

Based on the three clusters discovered through the modularity tool, we further subdivided the three modularity classes into six clusters for a more detailed topic-based interdisciplinary analysis. Table 6 summarizes the interdisciplinarity indices and centrality

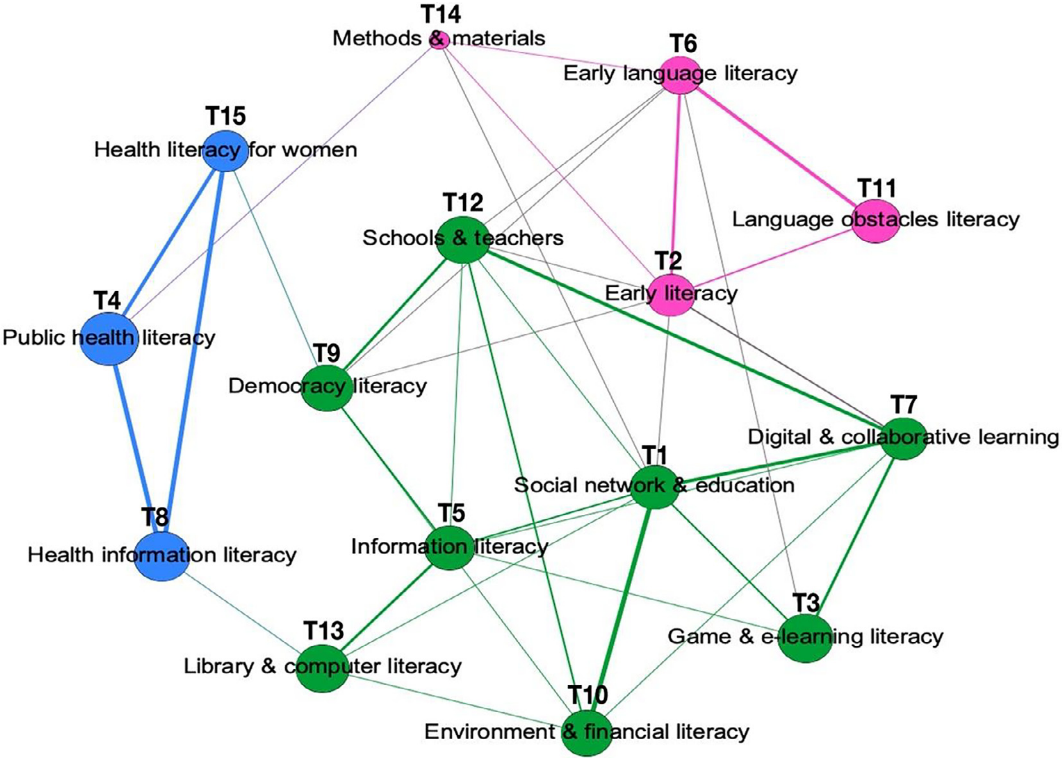

T14
Methods & materials
T6
Early language literacy
T15
Health literacy for women
T11
Language obstacles literacy
T12
Schools & teachers
T2
Early literacy
T4
Public health literacy
T9
Democracy literacy
T7
Digital & collaborative learning
T1
Social network & education
T5
Information literacy
T8
Health information literacy
T3
Game & e-learning literacy
T13
Library & computer literacy
T10
Environment & financial literacy

Fig. 5. A topic-based network visualized with Gephi software. (Node: topics; link: keyword co-occurrence; size: variety; color: modularity class).

values of topics by cluster. In this section, we comprehensively analyze the interdisciplinarity of each topic and the keyword-based relationship between topics based on the constructed network.

#1 Health literacy: T4, T8, T15

This cluster has a health theme detected by Gephi's modularity. Public health (T4) had both a large number and a balanced ratio of disciplines, whereas both overall keywords and highly ranked keywords were different and specialized among disciplines. Health information (T8) is linked to Library & computer literacy (T13), but overall the topics show low interdisciplinarity.

#2 Early literacy: T6, T2, T11

With the theme of early literacy, this cluster featured topics specializing in early language learning and language obstacles. Early language literacy (T6) has high interdisciplinarity with both overall keywords and highly ranked keywords similar to other academic fields. On the other hand, in Early literacy (T2), high-ranking keywords show a low interdisciplinarity that is not shared with other academic fields within a topic. They also exhibit a high betweenness centrality, so it can be seen that keywords are shared with other topics.

#3 Methods and materials: T14

T14 is an unbalanced topic with a small number of disciplines and a high distribution of several disciplines. Although most keywords are not similar between disciplines, the top-ranking keywords have a high interdisciplinarity that is intensively shared between disciplines. As a methodology, T14 is linked to the Social network & education (T1) and Early literacy (T2 and T6) clusters as well as the Health (T4) cluster. This implies the possibility of collaboration between social learning methods and educational materials in the fields of early literacy and health literacy.

#4 Social learning: T1, T3, T10

According to the results of the disparity index, Social network & education (T1) has high interdisciplinarity, where overall keywords are shared across academic fields. In particular, many keywords are shared with other topics due to high betweenness centrality. This cluster also includes methodologies such as social network analysis, so it is a topic with great potential for collaboration with other topics. It is connected to topics that contain specific themes of literacy, such as Game & e-learning literacy (T3) and Environment & financial literacy (T10). Topic trend analysis reveals a rising pattern in the late 2000s; thus, this is a research topic expected to actively collaborate with other academic fields.

#5 Information literacy: T5, T13

Information literacy (T5) exhibits high interdisciplinarity because highly ranked keywords have a diversity index value that is intensively shared with other academic fields. This topic has major connectivity with Library & computer literacy (T13), and also shares keywords with Social network & education (T1) and Democracy literacy (T9). T5 is a research topic that shows a rising pattern in the late 2000s in the topic distribution by year.

#6 Digital & collaboration: T7, T9, T12

Digital & collaborative learning (T7) has the same characteristics as T5, with keywords ranked highly according to the diversity index value similar to those of other academic fields. This topic has high connectivity with Schools & teachers (T12) and Social network & education (T1). Schools & teachers (T12) share many keywords with Democracy literacy (T9), which has a high betweenness centrality. T7 also exhibits a similar pattern to T5, showing a gradually rising topic trend in the late 2000s.

5. Discussion

This study presented a methodological framework for analyzing ***topic-based interdisciplinarity***. The proposed bottom-up approach effectively describes the relationships and structure of disciplines centered on particular research topics. There are three primary advantages to our methodology.

First, this paper provides an effective framework that can comprehensively consider the trends and relationships of topics and the characteristics of interdisciplinarity. Research topics are essential in finding ways to collaborate between disciplines. The present study understands interdisciplinarity as a feature of topic similarity within disciplines rather than as a feature of journals. Through this framework, research topics that can promote collaboration can be found in the current topic maps of academic fields, taking time and interaction characteristics into consideration.

Second, to overcome the predefined schemes, topics in this study were composed by applying the topic modeling technique. Furthermore, relationships were constructed as co-word networks based on topical similarities. As a result, our methodology represents the current dynamic and convergent knowledge system in a bottom-up manner. The application of the DMR topic technique enabled identification of the knowledge structure of current disciplines and, at the same time, made it possible to grasp the future direction of topic trends.

Third, we constructed a deep keyword model as a state-of-the-art technique to generate keywords with important semantic meanings that did not appear in publications. Previous studies have constructed corpora with a limited number of author-designed keywords and a ranking algorithm of extracted keywords. Moreover, rather than focusing on research domain, previous studies have constructed corpora by confining themselves to one or two small disciplines. In order to overcome these limitations, we have secured a large number of keywords that contain unrestricted disciplines and in-depth research topics established through a deep text mining technique. Our methodology makes it possible to construct an adequate corpus to analyze interdisciplinarity, which refers to the relationships between disciplines.

6. Conclusions

This study presented a comprehensive framework that can enable topic-based interdisciplinarity analysis for a specific research domain consisting of multiple academic fields through a bottom-up approach. Through topic modeling, interdisciplinarity, and network analysis in the literacy research domain, we discovered specific topics with high potential for collaboration between a number of disciplines.

The study's main contributions are as follows. First, its corpus is composed of massive data from 296 disciplines that are not defined as one or several disciplines. Through a deep-learning model, we obtained newly generated keywords considering semantic meaning from abstracts of 346,387 articles published from 1917 to 2021. This vast amount of data has enabled rich interpretation of borderless interactions and relationships between disciplines. Second, the study's DMR topic modeling was able to not only effectively detect the topics of a vast corpus but also predict the trends of topics through analysis of rising and falling patterns since the 2000s. Third, a framework for the integrated analysis of interdisciplinarity measurement results with topic modeling results was presented via network analysis. Through the four interdisciplinarity indices, we not only measured the similarity of keywords shared by disciplines within each topic but also analyzed the relationships between topics based on keyword co-occurrence.

Despite its contributions, the present study has some limitations. First, as the training data for a deep-learning model, the KP20K dataset was primarily acquired from the field of computer science. However, the collected keywords, which served as the study's data for analysis, were obtained from publications in a wide variety of disciplines. Therefore, it is necessary to improve the deep keyword model by training with more comprehensive discipline data to generate keywords while avoiding any form of bias and improving the research quality. Second, when keywords were mapped to disciplines to measure interdisciplinarity, keywords' semantic importance was not reflected in the calculation of the four indices. Considering keywords' semantic weights and converting keywords to disciplines would enable the measurement of more topical interdisciplinarity.

Two potential directions can be suggested for future research. First, it is necessary to demonstrate that our bottom-up methodology can uncover more research topics, relationships, and interdisciplinary characteristics in different datasets of other topics related to more unknown or recently established academic fields. Comparing our corpus, which is semantically rich and numerically vast, with those of prior studies might lead to a deeper discussion of corpus conditions for interdisciplinarity studies. Second, in addition to the four interdisciplinarity indices used in this study, it might be useful to apply more indices that measure diversity from other perspectives. This would further our understanding of interactions between disciplines from a wider perspective in a topic-based knowledge structure.

Topic-based interdisciplinarity analysis is different from citation analysis, which is a top-down approach (Xu et al., 2016). Topic-based interdisciplinarity analysis can reveal the characteristics of topics embedded in the literature spread across a vast number of disciplines and analyze relationships between topics. Our proposed framework is not limited to disciplines but serves as a guide to uncover the characteristics of topics and relationships between topics that are actively discussed in a research domain with high interdisciplinarity (e.g., literacy). As the boundaries between disciplines gradually blur and converge, interdisciplinary collaborative research has been increasing. Therefore, this study is significant by presenting a methodology that reveals topics with high collaboration potential and their relationships using keywords in over 200 disciplines. The simultaneous use of topic-based analysis and journal citation analysis is likely to better reveal interdisciplinary topics; thus, we suggest this as a direction for future research.

CRediT authorship contribution statement

Hyeyoung Kim: Formal analysis, Data curation, Conceptualization, Methodology, Writing – original draft. **Hyelin Park:** Formal analysis, Data curation, Conceptualization, Methodology, Writing – original draft. **Min Song:** Formal analysis, Data curation, Conceptualization, Methodology, Writing – original draft.

Acknowledgements

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2019S1A5C2A03083499). This work was also partly supported by the Yonsei University Research Grant of 2021.

Appendix A

Appendix A
Summary of ASJC classification.

Domain	Field	No. of Subfields
Physical Sciences	Chemical Engineering	115
	Chemistry	
	Computer Science	
	Earth and Planetary Sciences	
	Energy	
	Engineering	
	Environmental Science	
	Material Science	
	Mathematics	
	Physics and Astronomy	
	Multidisciplinary	
Health Sciences	Medicine	102
	Nursing	
	Veterinary	
	Dentistry	
	Health Professions	
	Multidisciplinary	
Social Sciences	Arts and Humanities	65
	Business, Management and Accounting	
	Decision Sciences	
	Economics, Econometrics and Finance	
	Psychology	
	Social Sciences	
	Multidisciplinary	
Life Sciences	Agricultural and Biological Sciences	51
	Biochemistry, Genetics and Molecular Biology	
	Immunology and Microbiology	
	Neuroscience	
	Pharmacology, Toxicology and Pharmaceutics	

The All Science Journal Classification Codes (ASJC) of the Scopus follows the in-depth scheme of disciplines (domain, field, and subfield). We selected ‘subfield’ to convert keywords into discipline information.

Appendix B

Appendix B
The removed top 40 high-frequency words.

‘education’, ‘literacy’, ‘children’, ‘learning’, ‘reading’, ‘teaching’, ‘improving_classroom_teaching’, ‘gender’, ‘health_literacy’, ‘learning_strategies’, ‘information_literacy’, ‘curriculum’, ‘internet’, ‘social_media’, ‘bibliometrics’, ‘higher_education’, ‘secondary_education’, ‘teacher_education’, ‘writing’, ‘language’, ‘assessment’, ‘mental_health’, ‘pedagogical_issues’, ‘communication’, ‘technology’, ‘health’, ‘health_care’, ‘science’, ‘elementary_education’, ‘culture’, ‘narrative’, ‘school’, ‘knowledge’, ‘professional_development’, ‘mathematics’, ‘adolescents’, ‘reading_comprehension’, ‘training’, ‘survey’, ‘collaboration’

The top 40 high-frequency words were mostly general nouns rather than research topics; thus, they were eliminated because they did not help to differentiate between topics.

Appendix C

Appendix C
Centrality and modularity calculation by topics.

Topic	Label	DegreeCentrality	BetweennessCentrality	ClosenessCentrality	Modularity
T1	Social network & education	17.000	13.571	0.667	1.000
T2	Early literacy	11.000	11.033	0.636	2.000
T3	Game & e-learning literacy	7.000	1.282	0.519	1.000
T4	Public health literacy	10.000	3.400	0.452	0.000
T5	Information literacy	10.000	5.250	0.583	1.000
T6	Early language literacy	11.000	8.329	0.583	2.000
T7	Digital & collaborative learning	15.000	2.093	0.560	1.000
T8	Health information literacy	11.000	2.833	0.438	0.000
T9	Democracy literacy	9.000	13.441	0.636	1.000
T10	Environment & financial literacy	10.000	2.889	0.560	1.000
T11	Language obstacles literacy	6.000	0.000	0.424	2.000
T12	Schools & teachers	13.000	3.986	0.609	1.000
T13	Library & computer literacy	6.000	7.317	0.538	1.000
T14	Methods & materials	4.000	9.178	0.583	2.000
T15	Health literacy for women	10.000	4.400	0.483	0.000

A topic-based network was constructed based on keyword co-occurrence, and the relationships between topics were analyzed using centrality values and modularity values.

References

Alzaidy, R., Caragea, C., & Giles, C. L. (2019). Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents. In *The world wide web conference* (pp. 2551–2557). 10.1145/3308558.3313642.

Asubiaro, T. V., & Badmus, O. M. (2020). Collaboration clusters, interdisciplinarity, scope and subject classification of library and information science research from Africa: An analysis of Web of Science publications from 1996 to 2015. *Journal of Librarianship and Information Science, 52*(4), 1169–1185. 10.1177/0961000620907958.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Third international AAAI conference on weblogs and social media*.

Billington, C. (2016). *How Digital Technology Can Support Early Language and Literacy Outcomes in Early Years Settings: A Review of the Literature*. National Literacy Trust: London, UK.

Blondel, V. D., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics, 2008*(10), P10008–P10012. 10.1088/1742-5468/2008/10/P10008.

Boechler, P., Dragon, K., & Wasniewski, E. (2014). Digital literacy concepts and definitions: Implications for educational assessment and practice. *International Journal of Digital Literacy and Digital Competence, 5*(4), 1–18. 10.4018/ijdldc.2014100101.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *Ucinet for Windows: Software for Social Network Analysis*. Harvard, MA: Analytic Technologies.

Bröder, J., Okan, O., Bauer, U., Bruland, D., Schlupp, S., Bollweg, T. M., et al. (2017). Health literacy in childhood and youth: A systematic review of definitions and models. *BMC Public Health, 17*(1), 1–25. 10.1186/s12889-017-4267-y.

Bu, Y., Li, M., Gu, W., & Huang, W. B. (2020). Topic diversity: A discipline scheme-free diversity measurement for journals. *Journal of the Association for Information Science and Technology, 72*(5), 523–539. 10.1002/asi.24433.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Friedkin, N. E. (1991). Theoretical foundations for centrality measures. *American Journal of Sociology, 96*, 1478–1504. 10.1086/229694.

García-Vallejo, D., Alcayde, A., López-Martínez, J., & Montoya, F. G. (2019). Detection of communities within the multibody system dynamics network and analysis of their relations. *Symmetry, 11*(12), 1525. 10.3390/sym11121525.

Gu, J., Lu, Z., Li, H., & Li, V. O. (2016). Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*.

Hicks, C. C., Fitzsimmons, C., & Polunin, N. V. C. (2010). Interdisciplinarity in the environmental sciences: Barriers and frontiers. *Environmental Conservation, 37*(4), 464–477. 10.1017/S0376892910000822.

Hu, J., & Zhang, Y. (2018). Measuring the interdisciplinarity of big data research: A longitudinal study. *Online Information Review, 42*(5), 681–696. 10.1108/OIR-12-2016-0361.

Leinster, T., & Cobbold, C. A. (2012). Measuring diversity: The importance of species similarity. *Ecology, 93*(3), 477–489. 10.1890/10-2402.1.

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology, 58*(9), 1303–1319. 10.1002/asi.20614.

Leydesdorff, L., & Goldstone, R. L. (2014). Interdisciplinarity at the journal and specialty level: The changing knowledge bases of the journal cognitive science. *Journal of the Association for Information Science and Technology, 65*(1), 164–177. 10.1002/asi.22953.

Leydesdorff, L., & Ivanova, I. (2020). The measurement of “interdisciplinarity” and “synergy” in scientific and extra-scientific collaborations. *Journal of the Association for Information Science and Technology, 72*(1), 387–402. 10.1002/asi.24416.

Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2018). Betweenness and diversity in journal citation networks as measures of interdisciplinarity: A tribute to Eugene Garfield. *Scientometrics, 114*(2), 567–592. 10.1007/s11192-017-2528-2.

Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019). Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. *Journal of Informetrics, 13*(1), 255–269. 10.1016/j.joi.2018.12.006.

Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., & Chi, Y. (2017). Deep keyphrase generation. *55th Annual Meeting of Association for Computational Linguistics* arXiv preprint arXiv:1704.06879 .

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404–411).

Mimno, D. M., & McCallum, A. (2008). Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. *In UAI, 24*, 411–418.

National Academies Committee on Facilitating Interdisciplinary Research. (2005). Committee on Science, Engineering and Public Policy (COSEPUP). *Facilitating Interdisciplinary Research*. Washington, DC: National Academies Press.

Owusu-Ansah, E. K. (2005). Debating definitions of information literacy: Enough is enough!. *Library Review, 54*(6), 366–374. 10.1108/00242530510605494.

Rafols, I., & Meyer, M. (2007). Diversity measures and network centralities as indicators of interdisciplinarity: Case studies in bionanoscience. *Proceedings of ISSI, 2*, 631–637.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics, 82*(2), 263–287. 10.1007/s11192-009-0041-y.

Rao, C. R. (1982). Diversity: Its measurement, decomposition, apportionment, and analysis. *Sankhya. Series A, 44*(1), 1–22.

Rousseau, R. (2018). The repeat rate: From Hirschman to Stirling. *Scientometrics, 116*(1), 645–653. 10.1007/s11192-018-2724-8.

Stirling, A. (1998). On the economics and analysis of diversity. *Science Policy Research Unit (SPRU), Electronic Working Papers Series, Paper, 28*, 1–156.

Stirling, A. (2007). A general framework for analysing diversity in science, technology, and society. *Journal of the Royal Society Interface, 4*(15), 707–719. 10.1098/rsif.2007.0213.

Stuart, A., Arnold, S., Ord, J. K., O’Hagan, A., & Forster, J. (1994). *Kendall’s Advanced Theory of Statistics*. New York: Wiley.

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology, 67*(10), 2464–2476. 10.1002/asi.23596.

Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS), 28*(4), 1–38. 10.1145/1852102.1852106.

Wolfe, A. W. (1997). Social network analysis: Methods and applications. *American Ethnologist, 24*, 136–137. 10.1525/ae.1997.24.1.219.

Xu, H., Guo, T., Yue, Z., Ru, L., & Fang, S. (2016). Interdisciplinary topics of information science: A study based on the terms interdisciplinarity index series. *Scientometrics, 106*(2), 583–601 https://doi-org-ssl.access.yonsei.ac.kr:8443/10.1007/s11192-015-1792-2.

Zhang, L., Janssens, F., Liang, L. M., & Glänzel, W. (2010). Journal crosscitation analysis for validation and improvement of journal-based subject classification in bibliometric research. *Scientometrics, 82*(3), 687–706. 10.1007/s11192-010-0180-1.

Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the Association for Information Science and Technology, 67*(5), 1257–1265. 10.1002/asi.23487.

Zhou, Q., Rousseau, R., Yang, L., Yue, T., & Yang, G. (2012). A general framework for describing diversity within systems and similarity between systems with applications in informetrics. *Scientometrics, 93*(3), 787–812. 10.1007/s11192-012-0767-9.