IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 7, 2021, accepted August 24, 2021, date of publication September 3, 2021, date of current version September 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3110111

Discovering Interdisciplinarily Spread Knowledge in the Academic Literature

MAIKO KAMADA, KIMITAKA ASATANI, MASARU ISONUMA, AND ICHIRO SAKATA, (Member, IEEE)
Graduate School of Engineering, The University of Tokyo, Tokyo 113-0033, Japan
Corresponding author: Kimitaka Asatani (asatani@ipr-ctr.t.u-tokyo.ac.jp)

This work was supported in part by the New Energy and Industrial Technology Development Organization (NEDO), Japan, under Grant P15009.

ABSTRACT With the increase in scientific publications and the diversification of research areas, science has become complex and interdisciplinary. Discovering important knowledge has become difficult even for researchers in specific domains. Previously proposed keyphrase extraction methods focus mainly on detecting intensively discussed topics in specific domains but do not distinguish concepts with the interdisciplinary spread from those discussed in narrow areas. Here, we propose a diffusion meme score that evaluates the knowledge diffusion distance in a paper citation network. The distance between papers that contain specific terms is measured by the network embedding space of the citation network. Using 57 million publication records from 48 years of Scopus, we evaluated newly appearing terms in and after 1975 in biomedical science papers using the proposed indicator. Approximately half of the top 20 terms were related to Nobel Prize or Clarivate Citation Laureates, and the top terms of the indicators were more likely to appear in Wikipedia than terms extracted using existing methods. Moreover, the top terms were unlikely to include specific minor diseases, which are often extracted using existing methods. Therefore, the diffusion meme score evaluates important terms from scientific literature and citation networks that are more interdisciplinary. Our method improves the understanding of young researchers regarding domains, the development of the history of science, and the evaluation of researcher contributions.

INDEX TERMS Citation network, interdisciplinary research, knowledge discovery, science of science.

I. INTRODUCTION

As science becomes massive and fragmented into specific domains [1], interdisciplinary research activities play an important role in developing scientific discoveries [2], [3]. Recent complex issues have become more important, such as climate change, food security, and healthy aging. In this context, funding agencies [4] and research organizations [5], [6] emphasize the interdisciplinarity of research. In specific domains, such as medical science, new therapies are developed by scientists in many disciplines, such as pharmacologists, molecular biologists, and neuroscientists, and involve scientists of public health, gerontology, zoology, education and engineering [7], [8]. For example, the hepatitis C virus was discovered to be the main cause of hepatitis by virologists [9]–[11], and treatment and prevention were established by scientists of internal medicine and preventive medicine [12]. Therefore, detecting such interdisciplinarily knowledge is important for gaining a better understanding of scientific achievements.

In general, learning a domain is not the result of understanding the fragments of knowledge. However, existing keyphrase extraction methods focus on detecting intensively discussed topics in specific domains [13], [14]. To better understand science, interdisciplinarily important knowledge that bridges academic fields should be extracted. Comprehending the connections between domains helps researchers who tend to make conservative choices to delve deeper into their areas of expertise [15] and find potential applications [16]. Moreover, this knowledge also helps science historians unravel the complicated evolutionary process of science and is essential for institutions that distribute budgets and award prizes.

Literature-based discovery [17], [18] is an intensive research domain for supporting researchers in finding exceptional concepts from the scientific literature. However, these

The associate editor coordinating the review of this manuscript and approving it for publication was Hocine Cherifi.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

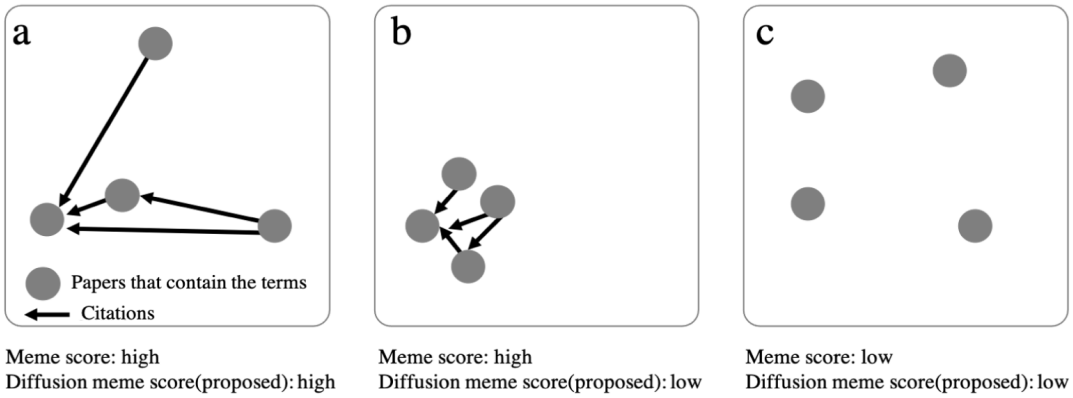


FIGURE 1. Differences in the forms of citation relationships that are highly valued by the meme score and diffusion meme score. A term that includes papers that cite each other and are published in distant domains is highly valued by the meme score and diffusion meme score (left figure). A term that includes papers that cite each other and are in similar domains is highly valued only by the meme score (central figure). A term used in some papers that do not cite each other is evaluated as zero in both indicators (right figure).

methods are highly domain-specific and are not applicable to large-scale datasets. Therefore, discovering exceptional interdisciplinary knowledge from a large data set is still a challenging issue [19]. In the context of natural language processing, discovering knowledge from the academic literature is approached as a keyphrase extraction task, which detects important terms from scientific publications [14], [20], [21]. Some unsupervised methods extract important terms by focusing on the appearance or relationship of terms [14]. For instance, TF-IDF [13] evaluated the appearance of terms in a document through arithmetic calculations of their frequencies in each document [22]. PositionRank [23] evaluated the importance of words by computing a PageRank-based score in a word co-occurrence network [24]. However, these indicators do not directly measure the interdisciplinary significance of scientific terms.

Chains of knowledge among scientists [25] should contain important information for evaluating interdisciplinary knowledge. Frequently, knowledge is represented as a term, and citation networks explicitly represent the chains of scientists' knowledge evaluations [26]. Therefore, for evaluating knowledge diffusion of a scientific term, it is straightforward to focus on a citation network among the papers that contain the term. Recently, Mao *et al*. detected important scientific terms in a domain that connect subfields by evaluating the amount of knowledge diffusion from the same domain [27]. However, these studies have focused only on a specific field comprising several thousand papers, and they use journal categories that do not effectively reflect the increasingly complex relationships among research disciplines. Closely related to our study, Kuhn *et al*. proposed the meme [28] score *Mm*, which evaluates the ''local'' importance of a term using a citation network [29]. The meme score evaluates the convergence of documents that include the term in the citation network and detects the knowledge that is discussed between scientists as a meme. Based on the idea of memes, the patterns of knowledge diffusion within and between disciplines have been examined [30], [31].

However, the meme score does not differentiate the terms used in a limited area (the middle of Fig. 1) or the terms that spread interdisciplinarily across academic fields (the left of Fig. 1). The former terms are likely to be highly specialized, such as the name of a minor surgery technique or minor proteins. Interdisciplinary concepts are assumed to be spread widely across academic fields because they are referenced in various fields. The distance between the endpoints of the citations indicates unanticipated knowledge transfer under the assumption that the distance between the fields of the two papers correlates with the similarity of the term occurrence of the papers. Therefore, an interdisciplinary concept can be evaluated by its spreading distance across academic fields. We use this idea to evaluate the unexpectedness of knowledge diffusion for detecting valuable interdisciplinary knowledge.

We propose a *diffusion meme score D*, which computes the sum of the distances of propagation across scientific fields and extracts exceptional scientific terms in terms of interdisciplinarity. Using this indicator, the terms that have propagated across diverse and distant communities obtain higher scores than those used by specific, narrow communities in a limited way. The distance indicates the difficulty or effort of the knowledge creation process. The accumulation of the distance describes each term's total impact on scientists in their research activity. Therefore, it makes sense that the top diffusion meme score terms are important terms (that scientists research frequently) in the process of academic evolution.

To measure the distances among academic fields, we obtain the latent representation of each node (paper) and measure the distance between the nodes. Conventionally, the reciprocal density of the citations between fields represents the distance between the fields. However, this indicator does not take into account the global structure of scientific fields. Recently, network representation learning

has been widely used in machine learning tasks because the obtained vector representation retains both local and global structures [32]. Thus, we measure the interdisciplinary uniqueness of terms by computing the diffusion distance with network representation learning (DeepWalk [33]). The detailed procedures are described in the materials and methods section.

By conducting a case study, we demonstrate that the diffusion meme score successfully discovers exceptional concepts in terms of interdisciplinarity. We calculated the diffusion meme scores of newly appearing terms in the text (title and abstract) and citations of 21 million biomedical science papers in and after 1975. We calculated the diffusion distances using 57 million papers in whole domains. Approximately half of the top 20 scientific terms in terms of diffusion meme score are related to notably exceptional concepts, including those that had won the Nobel Prize or Thomson Innovation Award. However, the (original) meme score provides a higher score for concepts that are researched intensively in a narrow area. This tendency is confirmed in an evaluation using outside databases. The scientific terms that are evaluated highly in terms of the diffusion meme score are more likely to appear in Wikipedia than those evaluated highly in terms of the meme score; however, the opposite result was observed in an evaluation using a disease database (MalaCards [34]) that contains many minor diseases. The proposed method demonstrates that the spread of a scientific term across diverse academic fields is strongly related to social evaluation. This insight is significant to clarify the process of academic development.

II. METHODS

A. DEFINITION OF THE DIFFUSION MEME SCORE

We define the natural logarithm of the sum of the distances d between the cited/citing references that contains the meme as the diffusion meme score D. 1 is added so that the logarithm can be calculated even if the sum of the distances is zero. We use network clustering to obtain small groups of articles that represent different research fields. Using the vector representation of the clusters obtained by embedding, the distance between the source clusters (cited) and destination (citing) documents is calculated as the Euclidean distance. The meme propagation between less involved clusters is large, and the propagation between closely related clusters is calculated to be small. Clustering and embedding are described in a later section.

Each term is evaluated using a citation network of papers that contained the term until ten years after the first 20 appearances of the term. The reason for the ten-year limit is to evaluate recent and old terms equally, supposing that most innovations are made in the decade after the terms become recognized.

$$D = \ln\Big(\sum_{e_s, e_t} d(e_s, e_t) + 1\Big) \tag{1}$$

$$d(e_s, e_t) = \cos(\vec{e_s}, \vec{e_t}) \tag{2}$$

d: Cosine distance between two nodes

e_s: Position of the cluster to which the cited reference belongs

e_t: Position of the cluster to which the citing reference belongs

B. CITATION NETWORK CLUSTERING

An unweighted directed network is constructed by connecting each literature node with an edge that connects the citations in the references from the source to the destination. We use the Leiden method [35] to classify the literature network into clusters of research fields. Clustering the literature citation network provides clusters of closely related references. Each of these clusters is defined as a research field. This categorization is used to obtain a distributed representation of a term or to calculate a term's distance.

The results of the clustering reflect the research activity of the citation relationship. This is the judgment of the authors of the article, who are familiar with the content, regarding the relevance of the content, and it allows for a better classification of research areas than methods based on journals and keywords.

After clustering approximately 57 million papers that have citation relationships with other papers, the clusters with more than 1000 references were further clustered twice (for a total of three recursive clustering steps), resulting in 24,908 clusters.

C. CLUSTER EMBEDDING

Graph embedding is a technique of projecting nodes in a graph onto a vector space. The vector representation of clusters indicates whether the clusters are in relatively similar or completely different fields. Several graph embedding methods have been proposed in the literature. In these methods, a long distance in the embedding space indicates the rarity of citations between them. It is difficult to choose the most adequate method by considering their embedding mechanisms. Thus, we evaluate several embedding methods and confirm the small difference between them (in ''Results - Comparison with other embedding methods''). In this paper, we adopt a method called DeepWalk [33].

The citation network is rewritten in the form of an inter-cluster citation network using the clustering results. Next, we randomly walk a certain number of times starting from a randomly selected node and obtain the series data of the visited nodes. Then, using Skip-Gram [36], we learn the variance representation vectors of each cluster to predict which clusters will appear around a given cluster. By projecting the citation network onto a high-dimensional (128-dimensional) space, the global structure of the citation network is preserved in the vector of each node while maintaining the local structure.

D. EXISTING METHODS

A variety of methods have been used to measure the information value of terms [18], [37]. Among them, the meme score M_m defined by Kuhn *et al.* [29] is a method that solves a

problem with the conventional methods, as it does not require a threshold of the number of times a term appears in the literature or expertise. The meme score M_m is determined by the frequency of the occurrence of a term m in the literature, f_m, and the heritability of the term m in the citation network, P_m.

$$M_m = f_m P_m \tag{3}$$

$$P_m = \frac{d_{m\to m}}{d_{\to m}+\delta} \Big/ \frac{d_{m\to \not m}+\delta}{d_{\to \not m}+\delta} \tag{4}$$

In (4), $d_{m\to m}$ indicates the number of publications such that the given meme term appears in the publication and in its cited publications. $d_{\to m}$ indicates the number of publications that cite publications that contain the meme term. $d_{m\to \not m}$ indicates the number of publications such that the given meme term appears in the publication and it does not cite publications that contain the meme term. $d_{\to \not m}$ indicates the number of publications that do not cite publications that contain the meme term. *delta* is a controlled noise for preventing the high evaluation for low frequent terms.

However, this method does not consider the distance of term propagation. It is suitable for evaluating terms that are closely discussed in a narrow community of knowledge and have been established in a research field, but it is not possible to evaluate terms that have interdisciplinary influences in many fields. In this paper, we propose a novel index that can evaluate influential terms that have an interdisciplinary spread in many fields.

E. RESULT EVALUATION

The evaluation of the proposed method was based on the two axes of the expertise and breadth of the extracted terms.

MalaCards [34], an exhaustive database of disease names, was used to assess whether the diffusion meme score can extract specialty terms. The number of diseases in this list among the terms extracted by the proposed method was treated as the extraction accuracy of the expertise terms. MalaCards integrates disease names and annotations for human diseases from 75 data sources. The diseases are assigned to 18 categories representing body regions such as blood, bone, immune system, and muscle and to six global categories such as cancer, genetics, and infections. As of January 20, 2020, MalaCards contained 22,371 diseases.

Wikipedia was used to verify whether the proposed method could extract a broad range of well-established terms in a global society. Wikipedia is an Internet encyclopedia operated by the Wikimedia Foundation and contains more than six million pages related to its content. Although the reliability of Wikipedia's content has been discussed frequently over many years [38]–[41], it was used as an index for evaluating the breadth of information because it is a useful tool for obtaining general information. Terms consisting of letters, numbers, and "−" from the list of titles were collected and downloaded from the Wikipedia database on January 1, 2020. The result was a total of 2,912,156 words. The percentage of the terms on Wikipedia among the terms extracted by the proposed method was treated as the overall accuracy of the terms' extraction.

III. DATA

Our analysis relied on 57,757,843 publication records from Scopus published between January 1970 and December 2018. We extracted terms that were formed of 3 words or fewer from 21,242,007 publication records that belonged to the category of Medicine and Immunology as well as Microbiology and focused primarily on titles and abstracts. We did not analyze terms from the first five years of the data, and we focused on literature published in and after 1975. Due to the large number of terms, the analysis was limited to terms that appeared more than 50 times in the decade beginning with the year in which the term appeared more than 20 times in total. Additionally, in this paper, terms that did not contain alphabetical letters and terms that included the year of publication were omitted as meaningless. As a result, 276,901 terms were analyzed, and each term was evaluated using a citation network of papers that contained the term from the first appearance of the term to 10 years after the first 20 appearances of the term.

IV. RESULTS

A. TERMS RATED HIGHLY BY THE DIFFUSION MEME SCORE

In preparation for the calculation of the diffusion meme score (1), we created a map of 57 million scientific papers. We applied network clustering recursively for 57 million publication records in the whole academic field of Scopus. The top 15 largest clusters are listed in Table.1. We found that the academic literature is composed of a divergent field of research. The clusters with more than 1,000 references were further clustered twice (for a total of three times), resulting in 24,908 clusters. Hereafter, we call a third-level cluster a cluster. We constructed a directed and weighted network of clusters, where each edge in the network represents the number of citations between the papers belonging to both end clusters. We calculated each cluster's 128-dimensional embedding from the network with DeepWalk. These clusters were visualized in 2D space by using t-distributed stochastic neighbor embedding (T-SNE) for dimension reduction (Fig. 2). We found that the clusters that belonged to the same top-level clusters were gathered in a certain space. The distance from the source to the destination of the term propagation via a citation was calculated in the 128-dimensional space. Formally, the distance $d(e_s, e_t)$ is the cosine distance between the positions of both clusters e_s, e_t to which the source and target nodes of the citations belong.

We calculated the diffusion meme scores (1) for words that newly appeared in and after 1975 in 21 million biomedical papers. We calculated the meme score of each term using the citations published within ten years of the first 20 appearances of the term. The purpose of the ten-year limit is to not overemphasize old terms. Table 2 shows the 20 terms that were rated most highly in the diffusion meme score D, the existing meme score Mm, and frequency (the

TABLE 1. Cluster size and label of the top 15 largest clusters.

cluster	size	Label	Frequent keywords
0	8,292,009	General	Education, China, Social, Monetary Policy, Economic System, Study
1	5,564,222	Material Science	Microstructure, Mechanical properties, Magnetic, Carbon Nanotube, Films
2	5,427,069	Informatics	Optimization, Simulation, Neural Networks, Algorithm, Linear Dynamics
3	3,866,952	Life-style related diseases	Obesity, Hypertension, Patients, Blood, Coronary Artery, Cardiovascular
4	3,703,869	Cancer	Brest Cancer, Apotosis, Cancer, Carcinoma, Radiotherapy, Tumor, Gene
5	3,559,961	Biology	Taxonomy, New species, Genetic, Soil Fauna, Nitrogen, Forest, Diversity
6	3,561,520	Intractable diseases	Alzheimer, Schizophenia, Neurons, Receptor, Brain, Rat, Antifreeze Protein
7	2,944,880	Cell Biology	Appotosis, Asthma, Expression, Chronic, Antibacterial, Virus, Immune
8	2,947,362	Structural Chemistry	Crystal Structure, Synthesis, Derivatives, Molecular, Spectroscopy, Metal
9	2,422,346	Environment	Remote sensing, Climate change, Water, Soil, Surface, Seismic Imaging
10	2,122,816	Analytical Chemistry	Chromatography, Spectrometry, Liquid, Determination, Marker Compounds
11	2,017,149	Thermodynamics	Heat, Model, Simulation, Numerical, Concrete, Finite Element Analysis
12	1,744,145	Quantum Optics	Plasma, Magnetic Field, Electron, Solar System, Neutron, Nuclear, Ion-drift
13	1,531,954	Infectious diseases	Escherichia, Tuberculosis, Infection, Coli Bacilli, Bacterial, Resistance
14	1,038,243	Obstetrics	Pregnancy, Sperm, Fetal, Ovarian Hyperstimulation Syndrome, In-Vitro

TABLE 2. Terms that are rated highly in each method. For the same term with different divisions or different names, the highest score is adopted (e.g., helicobacter, helicobacter pylori, and h. pylori). A term with + indicates that the person who discovered the drug or other substance related to the term received the Nobel Prize. Terms marked with * denote terms for which the discoverer was awarded the Clarivate Citation Laureates.

top terms rated high in D	D	the top terms rated high in M_m	M_m	the top frequent terms	$termfrequency$
helicobacter pylori+	10.870	endothelial keratoplasty	2.509	biomed central ltd	73,513
biomed central ltd	10.576	sialendoscopy	2.119	elsevier ireland ltd	65,589
elsevier ireland	10.554	onychomatricoma	1.992	springer science+business media	55,613
hepatitis c virus+	10.530	igg4-related	1.691	elsevier masson sas	36,002
toll-like+	10.509	sugammadex	1.604	lippincott williams	33,212
bevacizumab	10.424	transcranial direct current	1.396	informa healthcare	26,190
vascular endothelial growth factor*	10.344	cardiac contractility modulation	1.367	wolters kluwer health	24,909
IL-12	10.248	stump appendicitis	1.264	elsevier science b.v.	19,428
micrornas*	10.155	congenital pouch colon	1.235	elsevier inc	16,456
hiv-1+	10.061	purple urine bag syndrome	1.229	hiv infection+	16,306
chemokine	10.039	podoconiosis	1.153	journal compilation	15,995
cox-2	10.016	phosphaturic mesenchymal	1.137	main outcome measures	15,298
nitric oxide synthase+	10.006	genus nocardioides	1.136	helicobacter pylori+	14,231
leptin*	9.903	dicarbollylcobaltate	1.135	trial registration	13,820
infliximab	9.902	mpfl reconstruction	1.126	real-time pcr+	13,815
rituximab	9.901	nipple-sparing mastectomy	1.123	micrornas*	12,819
adiponectin	9.802	femoroacetabular impingement	1.106	pubmed	12,686
bcl-2	9.671	ebus-tbna	1.105	clinicaltrials	12,649
imatinib*	9.662	reduction malarplasty	1.083	study design	11,598
foxp3*	9.657	automated endothelial	1.079	elsevier gmbh	11,494

top 100 terms of each method are listed in Supplemental Table S2-4). Five terms (marked with +) are related to drugs or other substances that are the main contributions of scientists who received the Nobel Prize. The other five terms, marked with *, are terms for which the discoverer was awarded the Clarivate Citation Laureates.[1]

The results indicate that the diffusion meme score can determine terms that are likely to be included in Nobel Prize wins. For instance, the term ''hepatitis c virus'' was also included in research eligible for the 2020 Nobel Prize, while the analyzed data are from before 2019. Other examples are ''helicobacter'' and ''toll-like'', which were terms used in the research that won the Nobel Prize in Physiology or Medicine in 2005 and 2011. Although the paper citation network in 2018 is used for embedding clusters, the period of evaluated term-related citation networks is from 1992 to 2001 (former) and from 2001 to 2010 (latter). This indicates the possibility of predicting future Nobel Prizes related to these terms. Therefore, the proposed method can be used as an index to evaluate terms that are associated with

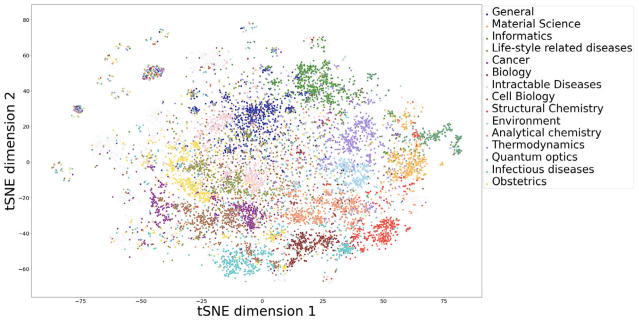


FIGURE 2. Visualization of clusters: The position of a cluster is a vector representation of 128 dimensions obtained by embedding and compressed into 2 dimensions by t-SNE. The clusters with more than 1,000 references were further clustered twice, resulting in 24,908 (sub-sub)clusters. Clusters belonging to the 15 largest top-level clusters are colored.

[1] An award presented by Clarivate Analytics, which identifies researchers who are likely to win the Nobel Prize in the near future.

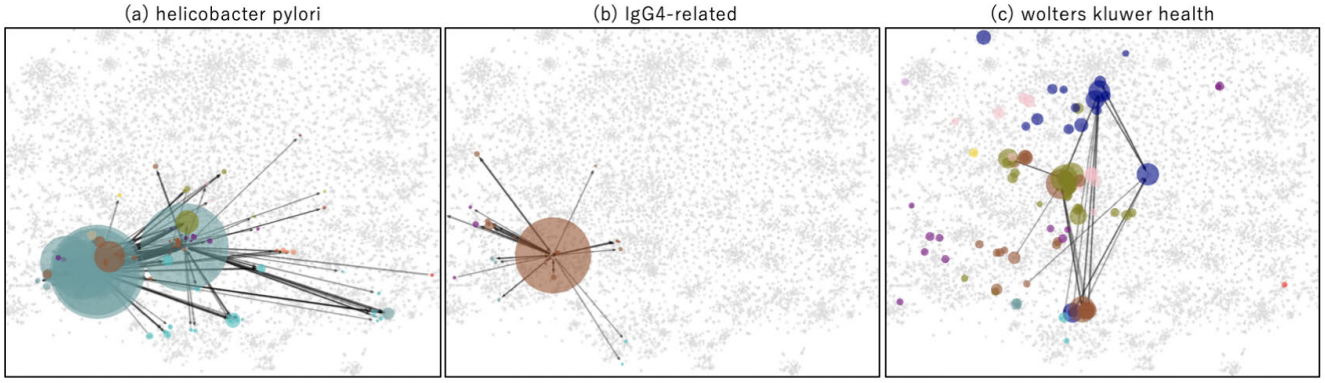


FIGURE 3. Diffusion network diagrams of a term that is highly rated in each method: The circles represent research field subclusters, and the color represents the main cluster to which each paper belongs. The size represents the number of times the term appears, and the edges indicate citations across disciplines for a term. The more citations a term has, the thicker the edges become. Terms that had high diffusion meme scores, such as (a) helicobacter pylori, have many research field clusters and edges between them. Terms that had high meme scores, such as (b) IgG4-related terms, appear in fewer research field clusters. Additionally, there is a hub cluster. Words that appear more frequently but are not highly rated in their respective meme scores, such as (c) Wolter Kluwer health, appear in diverse clusters but do not have many edges, indicating that they appear independently within clusters.

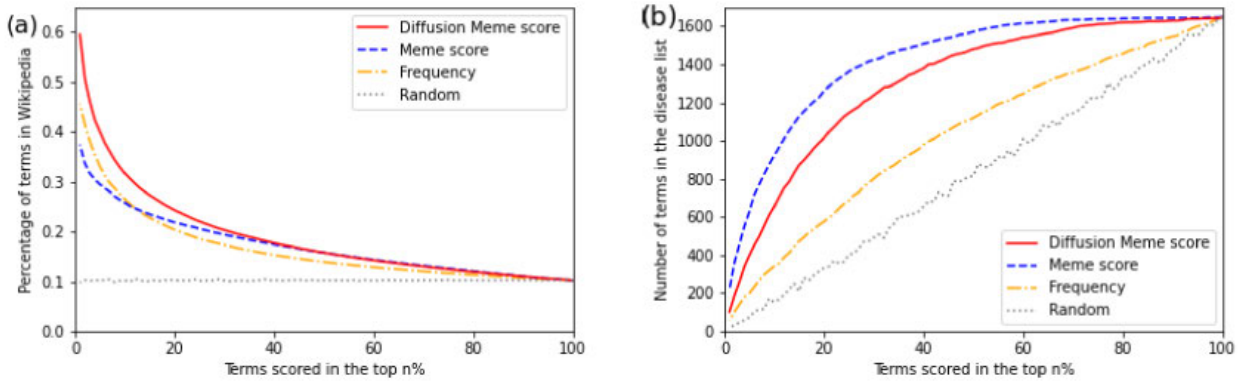


FIGURE 4. The evaluation of the diffusion meme score, meme score, and term frequency using Wikipedia and MalaCards: (a) Percentage of terms in Wikipedia among the terms extracted using each method illustrates that the diffusion meme score can extract terms that are more socially prevalent than other methods. (b) The number of terms in the disease list among the terms extracted using each method illustrates that the meme score can extract more specialized terms than other methods.

discoveries that have a great impact on society. However, diffusion meme scores falsely detect ''biomed central ltd'' because many papers in diverse fields contain this term and eventually connect to each other. This paper does not use the dictionary-based approach to remove such terms and evaluate the method correctly. However, the top 100 terms of each method in Supplemental Table S2 suggest that most misdetected words are related to publishers and are easy to remove using the dictionary-based approach.

However, the existing meme scores are high for medical devices, rare diseases, and surgical methods used for a small percentage of procedures. For instance, ''sialendoscopy'', a device used for salivary gland diseases, was highly rated by the meme score even though it was mentioned in only 280 papers. ''Onychomatricoma'', neoplastic nail lesions, was also highly rated, although it was mentioned in only 76 papers. This is consistent with the characteristics of the meme scores, which rate terms that are closely cited within a narrow field of study. The right side of Table 2, which lists the terms in order of their frequency of appearance, shows the publishers' names (''Biomed Central ltd'' and ''Elsevier Ireland ltd'') and general terms (''clinical trials'' and ''study design''). The diffusion meme score does not remove these words from the top lists completely but reduces their relative importance. The general terms are not evaluated highly in the diffusion and meme scores.

Fig. 3 illustrates the diffusion network diagrams of highly rated terms in each method. The terms that are highly rated in the diffusion meme score (the left of Fig. 3) have a large number of clusters, and most of the clusters have edges between them. Compared to the terms that are highly rated by the diffusion meme score, the terms that are rated higher in the meme score (the middle of Fig. 3) appear in fewer

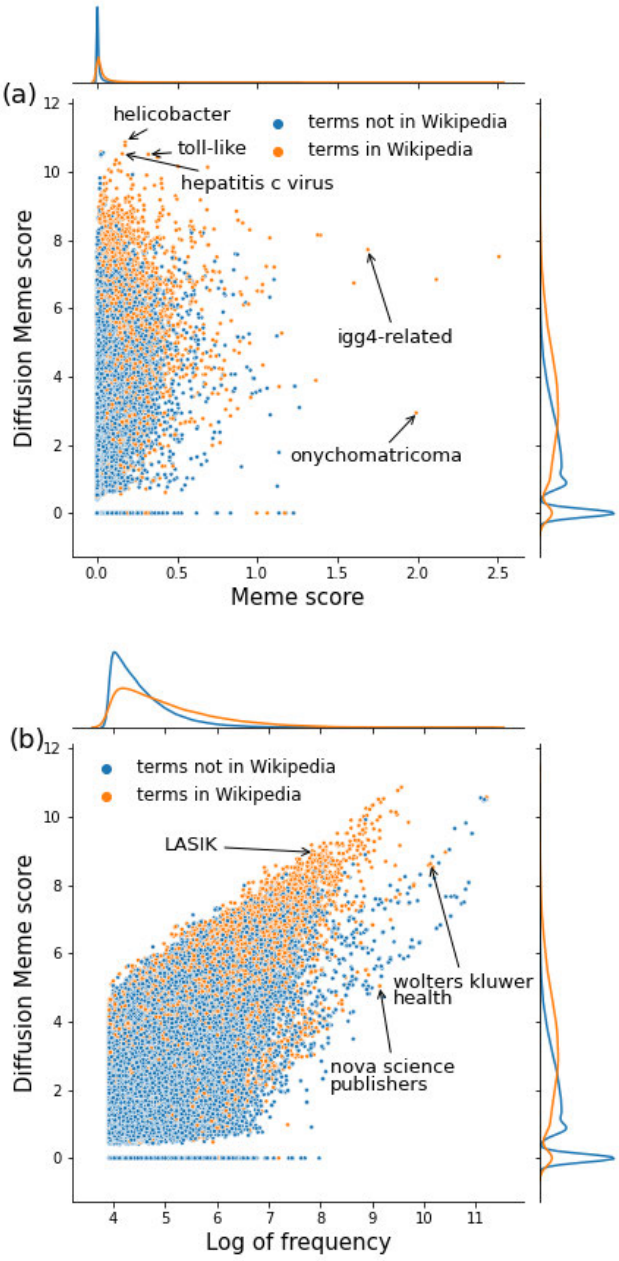


FIGURE 5. Differences in score distribution between diffusion meme score and other methods: (a) is a comparison of diffusion meme scores and meme scores, and (b) is a comparison of diffusion meme scores and the logarithm of the frequency of occurrence. Each dot represents a term, with Wikipedia words in orange and non-words in blue. The terms that scored highly on the meme scores or the logarithms of the frequency of occurrence are a mix of those listed on Wikipedia and those that are not. The distribution of listed and non-listed terms is split in the diffusion meme score.

clusters. As a representative example, IgG4-related[2] consists of one large node and a small connected node. This implies that the term propagation takes place mostly in one particular cluster. The right side of Fig. 3 illustrates the example of words (''wolters kluwer'') that are highly evaluated only by the word frequency. These words appear selectively in some specific clusters and can be evaluated highly in the TF-IDF of these clusters. This confirms that the original meme score can differentiate highly specialized terms from frequent terms.

B. STATISTICAL EVALUATION WITH WIKIPEDIA AND MalaCards

Wikipedia was used to assess whether the diffusion meme score could extract socially prevalent terms. The vertical axis in Fig. 4 illustrates the percentage of extracted terms that are listed on Wikipedia. In the case of random sampling, the vertical axis value hovers around 0.14, which is almost the same as the percentage of terms on Wikipedia in the total data used. The diffusion meme score demonstrates that around 60% of the top 10% of terms are listed on Wikipedia, and even when the number of extracted terms increases, the diffusion meme score contains the highest percentage of terms in Wikipedia compared with the other methods. The method using the number of appearances has the second-highest accuracy at first. This indicates that the diffusion meme score can evaluate terms related to generally important knowledge in society.

MalaCards [34], an exhaustive human disease database, was used to validate the extraction of expertise terms in the category of Medicine and Immunology as well as Microbiology. The right side of Fig. 4 indicates the number of terms in MalaCards that are in the top n% of each index. The vertical axis in the figure is the number of terms related to the disease's name among the extracted terms. The case where the terms are randomly extracted in addition to using the proposed diffusion meme score D and the comparison method, that is, the number of occurrences and the existing meme score Mm, is indicated in gray. This figure illustrates that the proposed method and comparison method always extract more terms than the random extraction case. Comparing each index, the results demonstrate that we were able to extract more disease name-related terms, always in the order of meme score, diffusion meme score, and the number of appearances.

The meme score is more likely to be evaluated when the propagation is tight and there is a community that actively discusses the term of interest. Therefore, in evaluating a list of disease names, including rare diseases, meme scores are useful for evaluating specialized terms discussed in minor communities. For example, the top 10% of terms in the existing meme score includes a genetic disease called Pallister-Killian [42], which appears in only 51 references, and an infectious disease called neuroschistosomiasis [43], a disease caused by a tropical worm, which appears in only 53 references. However, there are many terms that the meme score assesses that are confined to a narrow knowledge community; they are important within that community but are not widely known in society. In contrast, the diffusion meme score tends to be higher for terms that have spread to as many communities or distant communities as possible, which is why

[2]IgG4-related disease (IgG4-RD) is a chronic inflammatory condition characterized by tissue infiltration with lymphocytes and IgG4-secreting plasma cells, various degrees of fibrosis (scarring), and a usually prompt response to oral steroids. IgG4-RD has an incidence rate of 0.28-1.08 per 100,000 people.

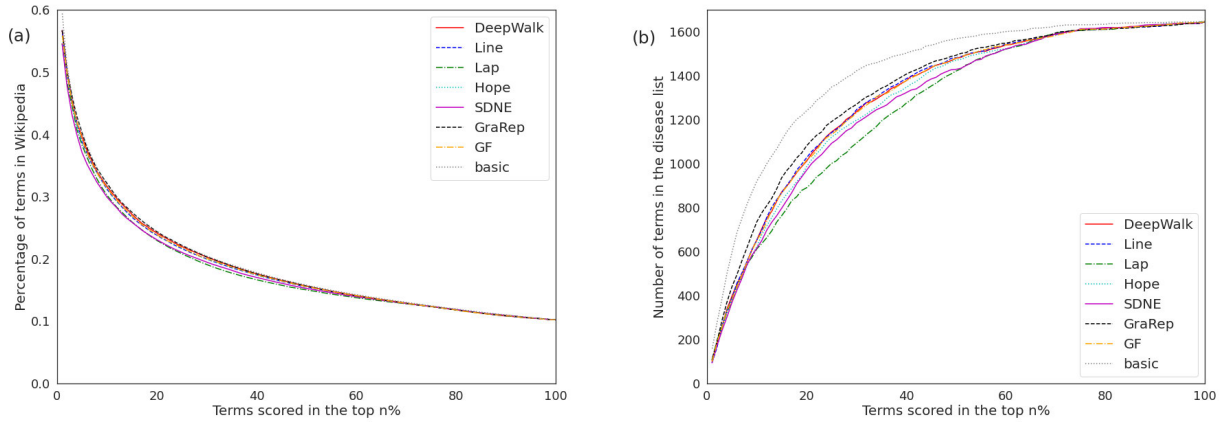


FIGURE 6. The difference in terms extracted using multiple embedding methods: (a) Comparison of the percentage of terms in Wikipedia among the terms extracted using multiple embedding methods. (b) The number of terms in the disease list among the terms extracted using multiple embedding methods: (a) illustrates that it does not make a significant difference which embedding method is used, and (b) illustrates that the basic method has the highest score for the extraction of highly specialized terms, showing a similar trend to existing meme methods.

the Wikipedia-based evaluation demonstrates better accuracy with the diffusion meme score. These analyses indicate that the meme score is adequate for comprehensively detecting topics discussed in academic papers and that the diffusion meme score is useful for detecting important interdisciplinary knowledge.

We also confirmed the differences among the diffusion meme score, meme score, and term frequency by the scatter plot of each term in Fig. 5. Most of the top words in both metrics appeared in Wikipedia. However, the figure illustrates that the results of these methods are not highly correlated. The terms ''helicobacter'' and ''onychomatricoma'' discussed above are plotted in the figure. The diffusion meme score evaluates words that are not evaluated in the meme score, and the diffusion meme score is more tightly correlated with term frequency (Fig. 5(b)). It is assumed that knowledge diffusion via citation networks is an essential process for a term (excluding publisher names) gaining high popularity. However, the difference in the top-evaluated words in both methods was significant. Typical examples include ''LASIK'' and ''nova science publishers'': the former is interdisciplinary applied surgery used to improve visual acuity, and the latter is the name of an American journal publisher.

C. COMPARISON WITH OTHER EMBEDDING METHODS

In this paper, DeepWalk was chosen as the embedding method to calculate the distance between clusters in (1). Here, we examine the differences in the results that appear when this embedding is replaced by other methods (Line [44], Laplacian eigenmaps (Lap) [45], HOPE [46], SDNE [47], GraRep [48], and graph factorization (GF) [49]). In addition, instead of cluster embedding using the citation relation, we set the basic method that calculates all diffusion distances as 1.

Fig. 6 shows how the accuracy changes when other embedding methods are used. In terms of broad and global importance, there was no significant difference between the methods. On the other hand, the basic method extracts the most words with high expertise, which is similar to the meme score. The basic method does not involve the distance between clusters and therefore cannot distinguish between interdisciplinary terms and terms that are narrowly discussed in closed clusters.

V. DISCUSSION

Our proposed method (the diffusion meme score) can extract terms from the corpus that are important globally. Most of the top 20 diffusion meme score terms are important in the medical science field, and some are Nobel prize-related terms. Most of the other top diffusion meme score terms are significantly important. For example, hepatitis c virus (HCV) has an interdisciplinary spread (71 million people are infected) and causes liver cancer [50]. Most of the top 200 terms are also important in medical science (such as Catenin, endothelin-1, and MCP-1). These terms provide a brief overview of recent progress in medical science. Our results confirm that the diffusion meme score, focusing on the spread of terms across disciplines, is effective, as it allows us to extract terms that impact global society.

The idea that the distance (surprise) of information spreading represents the importance of the information differs from other major term extraction methods such as TF-IDF [13] and the meme score [29]. The diffusion meme score applies to documents with relationship links, such as academic paper datasets of other domains and patent documents with reference data. Additionally, the diffusion meme score can explore important knowledge of human communications such as Twitter and other social networking sites data. Information diffusion on Twitter [51] has been intensively researched, especially for fake news [52]. Our approach may contribute to examining each path and estimating the global impact of information.

VI. CONCLUSION

We proposed the diffusion meme score D, which evaluates the knowledge diffusion distance in a paper citation network. The distance that is calculated in the network embedding space of the citation network indicates the difficulty or effort of the knowledge creation process. We confirmed that the sum of the distances indicates the importance of knowledge in science and society. Approximately half of the top 20 terms are related to Nobel Prize or Clarivate Citation Laureates, and the top terms of the indicators are more likely to appear in Wikipedia than terms extracted using existing methods. Our method improves the means by which young researchers can understand domains, the development of the history of science, and the evaluation of the contributions of researchers. The extracted important knowledge may provide a quick look at medical science for students, researchers and academic administrators. The diffusion meme score D is applicable to any other data composed of texts and their relationships.

The diffusion meme score does not evaluate the knowledge that is not represented as a term, and the correction of word polysemy and ambiguity leads to better results. However, this has a limited impact on the results because scientists tend to define concepts, objects, processes, and facts as terms and tend to use well-defined words. The ambiguity of a term between the cited and citing papers causes the decrease of the diffusion meme score. However, the effect is not significantly greater than that of other indicators, such as the term frequency and the original meme score. Another limitation is that documents without relationship data are not within the scope of the diffusion meme score. However, our method is applicable to these data using link prediction between documents, which is an important subject for data mining researchers. There is a possibility that implementing other distance calculation methods will improve the results of our method.

AUTHOR CONTRIBUTIONS STATEMENT

Kimitaka Asatani and Maiko Kamada designed the model and the computational framework. Maiko Kamada analyzed the data with the support of Kimitaka Asatani. Maiko Kamada and Kimitaka Asatani wrote the manuscript. Masaru Isonuma and Ichiro Sakata reviewed the study and assisted with the data analysis. Ichiro Sakata supervised the project.

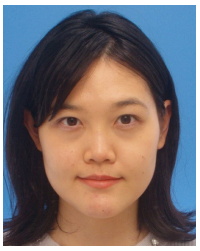
ACKNOWLEDGMENT

(Maiko Kamada and Kimitaka Asatani contributed equally to this work.)

REFERENCES

[1] S. Milojević, "Quantifying the cognitive extent of science," *J. Informetrics*, vol. 9, no. 4, pp. 962–973, Oct. 2015.
[2] M. Stember, "Advancing the social sciences through the interdisciplinary enterprise," *Social Sci. J.*, vol. 28, no. 1, pp. 1–14, Mar. 1991. [Online]. Available: http://www.sciencedirect.com/science/article/pii/036233199190040B
[3] J. T. Klein, "Evaluation of interdisciplinary and transdisciplinary research: A literature review," *Amer. J. Preventive Med.*, vol. 35, no. 2, pp. S116–S123, 2008.
[4] C. Lyall, A. Bruce, W. Marsden, and L. Meagher, "The role of funding agencies in creating interdisciplinary knowledge," *Sci. Public Policy*, vol. 40, no. 1, pp. 62–71, Feb. 2013.
[5] T. R. Cech and G. M. Rubin, "Nurturing interdisciplinary research," *Nature Struct. Mol. Biol.*, vol. 11, no. 12, pp. 1166–1169, Dec. 2004.
[6] M. Sugiyama, I. Sakata, H. Shiroyama, H. Yoshikawa, and T. Taniguchi, "Research management: Five years on from Fukushima," *Nature News*, vol. 531, no. 7592, p. 29, 2016.
[7] T. K. Woodruff, "Oncofertility: A grand collaboration between reproductive medicine and oncology," *Reproduction*, vol. 150, no. 3, pp. S1–S10, Sep. 2015.
[8] P. Varkey, S. P. Karlapudi, and K. E. Bennet, "Teaching quality improvement: A collaboration project between medicine and engineering," *Amer. J. Med. Qual.*, vol. 23, no. 4, pp. 296–301, Jul. 2008.
[9] A. A. Kolykhalov, E. V. Agapov, K. J. Blight, K. Mihalik, S. M. Feinstone, and C. M. Rice, "Transmission of hepatitis C by intrahepatic inoculation with transcribed RNA," *Science*, vol. 277, no. 5325, pp. 570–574, Jul. 1997.
[10] Q. Choo, G. Kuo, A. Weiner, L. Overby, D. Bradley, and M. Houghton, "Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome," *Science*, vol. 244, no. 4902, pp. 359–362, Apr. 1989.
[11] H. J. Alter, P. V. Holland, R. H. Purcell, J. J. Lander, S. M. Feinstone, A. G. Morrow, and P. J. Schmidt, "Posttransfusion hepatitis after exclusion of commercial and hepatitis-B antigen-positive donors," *Ann. Internal Med.*, vol. 77, no. 5, pp. 691–699, 1972.
[12] D. Lavanchy, "Evolving epidemiology of hepatitis C virus," *Clin. Microbiol. Infection*, vol. 17, no. 2, pp. 107–115, Feb. 2011.
[13] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Documentation*, vol. 60, no. 5, pp. 493–502, Oct. 2004.
[14] E. Papagiannopoulou and G. Tsoumakas, "A review of keyphrase extraction," *WIREs Data Mining Knowl. Discovery*, vol. 10, no. 2, Mar. 2020, Art. no. e01339.
[15] J. G. Foster, A. Rzhetsky, and J. A. Evans, "Tradition and innovation in scientists' research strategies," *Amer. Sociol. Rev.*, vol. 80, no. 5, pp. 875–908, 2015, doi: 10.1177/0003122415601618.
[16] M. Cokol, I. Iossifov, C. Weinreb, and A. Rzhetsky, "Emergent behavior of growing knowledge about molecular interactions," *Nature Biotechnol.*, vol. 23, no. 10, pp. 1243–1247, Oct. 2005.
[17] D. R. Swanson and N. R. Smalheiser, "An interactive system for finding complementary literatures: A stimulus to scientific discovery," *Artif. Intell.*, vol. 91, no. 2, pp. 183–203, 1997.
[18] J. Preiss, M. Stevenson, and R. Gaizauskas, "Exploring relation types for literature-based discovery," *J. Amer. Med. Inform. Assoc.*, vol. 22, no. 5, pp. 987–992, Sep. 2015.
[19] S. Henry and B. T. McInnes, "Literature based discovery: Models, methods, and trends," *J. Biomed. Inform.*, vol. 74, pp. 20–32, Oct. 2017.
[20] T. D. Nguyen and M.-Y. Kan, "Keyphrase extraction in scientific publications," in *Proc. Int. Conf. Asian Digit. Libraries*. Berlin, Germany: Springer, 2007, pp. 317–326.
[21] S. N. Kim, O. Medelyan, M.-Y. Kan, and T. Baldwin, "SemEval-2010 task 5: Automatic keyphrase extraction from scientific articles," in *Proc. 5th Int. Workshop Semantic Eval.*, 2010, pp. 21–26.
[22] A. Aizawa, "An information-theoretic perspective of TF–IDF measures," *Inf. Process. Manage.*, vol. 39, no. 1, pp. 45–65, 2003.
[23] C. Florescu and C. Caragea, "Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 1105–1115.
[24] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford InfoLab, Stanford, CA, USA, Tech. Rep., 1999.
[25] D. Shahaf, C. Guestrin, and E. Horvitz, "Metro maps of science," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2012, pp. 1122–1130.
[26] D. Jurgens, S. Kumar, R. Hoover, D. McFarland, and D. Jurafsky, "Measuring the evolution of a scientific field through citation frames," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 391–406, Dec. 2018.
[27] J. Mao, Z. Liang, Y. Cao, and G. Li, "Quantifying cross-disciplinary knowledge flow from the perspective of content: Introducing an approach based on knowledge memes," *J. Informetrics*, vol. 14, no. 4, Nov. 2020, Art. no. 101092.
[28] R. Dawkins, *The Selfish Gene*. Oxford, U.K.: Oxford Univ. Press, 1976.
[29] T. Kuhn, M. Perc, and D. Helbing, "Inheritance patterns in citation networks reveal scientific memes," *Phys. Rev. X*, vol. 4, no. 4, 2014, Art. no. 041036.

[30] Z. Liang, J. Mao, Y. Cao, and G. Li, "Idea diffusion patterns: SNA on knowledge meme cascade network," in *Proc. ISSI*, 2019, pp. 2612–2613.
[31] X. Sun and K. Ding, "Identifying and tracking scientific and technological knowledge memes from citation networks of publications and patents," *Scientometrics*, vol. 116, no. 3, pp. 1735–1748, 2018.
[32] D. Zhang, J. Yin, X. Zhu, and C. Zhang, "Network representation learning: A survey," *IEEE Trans. Big Data*, vol. 6, no. 1, pp. 3–28, Mar. 2020.
[33] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2014, pp. 701–710, doi: 10.1145/2623330.2623732.
[34] S. Espe, "MalaCards: The human disease database," *J. Med. Library Assoc.*, vol. 106, no. 1, p. 140, Jan. 2018.
[35] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: Guaranteeing well-connected communities," *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, Dec. 2019.
[36] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[37] O. Bodenreider, "The unified medical language system (UMLS): Integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, pp. D267–D270, Jan. 2004.
[38] R. Sumi, T. Yasseri, A. Rung, A. Kornai, and J. Kertesz, "Edit wars in Wikipedia," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust IEEE 3rd Int. Conf. Social Comput.*, Oct. 2011, pp. 724–727.
[39] T. Yasseri, R. Sumi, A. Rung, A. Kornai, and J. Kertész, "Dynamics of conflicts in Wikipedia," *PLoS ONE*, vol. 7, no. 6, Jun. 2012, Art. no. e38869.
[40] E. Borra, E. Weltevrede, P. Ciuccarelli, A. Kaltenbrunner, D. Laniado, G. Magni, M. Mauri, R. Rogers, and T. Venturini, "Societal controversies in Wikipedia articles," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, Apr. 2015, pp. 193–196.
[41] A. M. Wilson and G. E. Likens, "Content volatility of scientific topics in wikipedia: A cautionary tale," *PLoS ONE*, vol. 10, no. 8, Aug. 2015, Art. no. e0134454.
[42] A. Schinzel, "Tetrasomy 12p (Pallister-Killian syndrome)," *J. Med. Genet.*, vol. 28, no. 2, p. 122, 1991.
[43] F. J. Carod-Artal, "Neuroschistosomiasis," *Expert Rev. Anti-Infective Therapy*, vol. 8, no. 11, pp. 1307–1318, Nov. 2010, doi: 10.1586/eri.10.111.
[44] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "LINE: Large-scale information network embedding," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 1067–1077.
[45] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.
[46] M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu, "Asymmetric transitivity preserving graph embedding," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1105–1114.
[47] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1225–1234.
[48] S. Cao, W. Lu, and Q. Xu, "GraRep: Learning graph representations with global structural information," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, New York, NY, USA, Oct. 2015, pp. 891–900, doi: 10.1145/2806416.2806512.
[49] A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola, "Distributed large-scale natural graph factorization," in *Proc. 22nd Int. Conf. World Wide Web (WWW)*, 2013, pp. 37–48.
[50] T. W. H. Organization. (2020). *Hepatitis C*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hepatitis-c
[51] D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 695–704.
[52] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, May 2018.



MAIKO KAMADA was born in Kyoto, Japan, in 1997. She received the B.A. degree in social engineering from The University of Tokyo, Tokyo, Japan, in 2020, where she is currently pursuing the M.S. degree with the Graduate School of Engineering.

KIMITAKA ASATANI was born in Nishinomiya, Japan, in 1984. He received the B.A. degree in material engineering and the M.A. degree in computer science, and the Ph.D. degree in computer science from The University of Tokyo, in 2015. He is currently an Assistant Professor (Project) with The University of Tokyo. He studies complex networks and their applications, focusing particularly on social and knowledge relationships in scientific articles and social networking services.

MASARU ISONUMA was born in Tokyo, Japan, in 1992. He received the B.S. degree in engineering and the M.S. degree in engineering from The University of Tokyo, Tokyo, in 2015 and 2017, respectively, where he is currently pursuing the Ph.D. degree with the Graduate School of Engineering. His research interests include document summarization and latent structure induction.

ICHIRO SAKATA (Member, IEEE) was born in Osaka, Japan, in 1966. He received the B.A. degree in economics from The University of Tokyo, Tokyo, Japan, in 1989, the M.A. degree in international economics and finance from Brandeis University, Boston, MA, USA, in 1997, and the Ph.D. degree in environmental and ocean engineering from The University of Tokyo, in 2003. He became a Full Professor, in 2008, after working with the Ministry of Economy, Trade and Industry. He has held several appointments at The University of Tokyo, including the positions of a special advisor to the President , from 2015 to 2019 the Director of the Policy Alternatives Research Institute , from 2014 to 2016 the Head of the Vision Formation Group, FSI, since 2016, and the Head of the Department of TMI, Graduate School of Engineering , from 2016 to 2017. He is currently a Special Advisor to the President and a Professor with the School of Engineering, The University of Tokyo. His research interests include innovation management, technological forecasting, strategic research planning, and computational social science. He proposed the concept of technology informatics. He is a member of the Engineering Academy of Japan. Regarding social contributions, he served as a Special Advisor for the Ministry of Health, Labor and Welfare and the Ministry of Reconstruction in Japan.