Chapter 11

Mapping Science Based on Research Content Similarity

Takahiro Kawamura, Katsutaro Watanabe,
Naoya Matsumoto and Shusaku Egami

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.77067

Abstract

Maps of science representing the structure of science help us understand science and technology development. Thus, research in scientometrics has developed techniques for analyzing research activities and for measuring their relationships; however, navigating the recent scientific landscape is still challenging, since conventional inter-citation and co-citation analysis has difficulty in applying to recently published articles and ongoing projects. Therefore, to characterize what is being attempted in the current scientific landscape, this article proposes a content-based method of locating research articles/projects in a multi-dimensional space using word/paragraph embedding. Specifically, for addressing an *unclustered* problem, we introduced cluster vectors based on the information entropies of technical concepts. The experimental results showed that our method formed a clustered map from approx. 300 k IEEE articles and NSF projects from 2012 to 2016. Finally, we confirmed that formation of specific research areas can be captured as changes in the network structure.

Keywords: map of science, content-based, paragraph vector, information entropy, clustering

1. Introduction

In 1965, Price [1] proposed studying science using scientific methods. Since then, research in scientometrics has developed techniques for analyzing research activities and for measuring their relationships and constructed maps of science, one of the major topics in scientometrics, that provides a bird's eye view of the scientific landscape. Maps of science have been useful tools for understanding the structure of science, their spread, and interconnection of disciplines. By knowing such information, science, and technology enterprises can

anticipate changes, especially those initiated in their immediate vicinity. Research laboratories and universities that are organized according to the established standards of disciplinary departments can understand an organization's environment. Furthermore, such maps are important to policy analysts and funding agencies. Since research funding should be based on quantitative and qualitative scientific metrics, they usually perform several analyses on the map with statistical analysis and careful examination by human experts. However, conventional approaches to understanding research activities focus on what authors told us about past accomplishments through inter-citation and co-citation analysis of published research articles. Thus, ongoing project and the recently published articles that do not have enough citations have not been analyzed.

Therefore, we propose to analyze them using a content-based method using natural language processing (NLP) techniques. Recently, word/paragraph embedding has been proposed for finding relationships between unstructured descriptions. Such embedding techniques represent words and paragraphs as real-valued vectors of several hundred dimensions. The distances between the descriptions are calculated from the similarities between vectors. Thus, we constructed a new mapping tool that represents the recent scientific trends, where nodes represent research projects or the articles that are linked by certain distances of the content similarity. Moreover, we drew a map from approx. 300,000 IEEE articles and National Science Foundation (NSF) projects, and then from its chronological changes we obtained some findings regarding the formation processes of research areas.

The remainder of this chapter is organized as follows. In Section 2 discusses related work, and Section 3 describes our proposed method for calculating the content similarity and its evaluations. Then, Section 4 introduces our tool, Mapping Science, and we confirm on the map the formation process of research areas such as the Internet of Things in Section 5, final conclusions and suggestions for future work are provided in Section 6.

2. Related work

Maps of Science (http://mapofscience.com/) are a well-known website. Katy et al. also provides Sci2Tool visualization tools [2] and maps of journals and documents [3]. In Japan, National Institute of Science and Technology Policy (NISTEP) provides Science Map (http://www.nistep.go.jp/wp/wp-content/uploads/ScienceMapWebEdition2014.html). In such studies, the similarity between journals and articles is calculated based on the cosine and/or Jaccard similarity of inter-citation and co-citation. These maps promote interdisciplinary research collaboration, but citation analysis cannot be utilized for ongoing projects and recently published articles, although project descriptions will eventually include articles in their research results.

Funding agencies and publishers generally have their own classification systems. Projects/articles have more than one code; thus, interdisciplinary projects can be found by searching multi-labeled projects. However, even if two projects/articles are assigned the same category, their similarity may not be found. Moreover, funding agencies and publishers use different categories, and there is no comprehensive scheme for characterizing projects or articles; thus, they cannot be compared between different agencies or publishers. For example, comparing

articles with Association for Computing Machinery classification (https://www.acm.org/publications/class-2012) with Springer Nature classification requires taxonomy exchanges.

Therefore, several content-based methods are proposed in the related literature. Previous studies have examined automatic topic classification using probabilistic latent semantic analysis (pLSA) [4] and latent Dirichlet allocation (LDA) [5]. One uses LDA to find the five most probable words for a topic, and each document is viewed as a mixture of topics [6]. This approach can classify documents across different agencies and publishers. However, the similarity between projects/articles cannot be computed directly. In this regard, the National Institutes of Health (NIH) Visual Browser [7, 8] (http://nihmaps.org/index.php) computed the similarities between projects as the mixture of classification probability to each topic based on pLSA, using the average symmetric Kullback-Leibler divergence function [9]. However, this similarity is a combination of probabilities; that is, it is not derived from sentence context. Other studies are also based on the similarity between sets of words (bag-of-word) included in documents like term frequency-inverse document frequency (TF-IDF), and not considering the sentence context.

By contrast, a word/paragraph vector, which is a distributed representation of words and paragraphs, is attracting attention in NLP. Assuming that context determines the meaning of a word [10], words appearing in similar contexts are considered to have a similar meaning. In the basic form, a word vector is represented as a matrix, whose elements are the co-occurrence frequencies between a word w with a certain usage frequency in the corpus and words within a fixed window size c from w. A popular representation of word vectors is word2vec [11, 12]. Word2vec creates word vectors using a two-layered neural network obtained by a skip-gram model with negative sampling. Specifically, word vectors are obtained by calculating the maximum likelihood of objective function L in Eq. (1), where T is the number of words with a certain usage frequency in the corpus. Word2vec clusters words with similar meanings in a vector space.

$$L = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p\left(w_{t+j} \mid w_t\right) \tag{1}$$

In addition, Le and Mikolov [13] proposed a paragraph vector that learns fixed-length feature representations using a two-layered neural network from variable-length pieces of texts such as sentences, paragraphs, and documents. A paragraph vector is considered another word in a paragraph and is shared across all contexts generated from the same paragraph but not across paragraphs. The contexts are fixed length and sampled from a sliding window over the paragraph. The paragraph vectors are computed by fixing the word vectors and training the new paragraph vector until convergence, as shown in Eq. (2).

$$L = \sum_{t=1}^{T}\log p\left(w_t \mid w_{t-c}, \ldots, w_{t+c}, d_i\right) \tag{2}$$

where d_i is a vector for a paragraph i that includes w_t. Whereas word vectors are shared across paragraphs, paragraph vectors are unique among paragraphs and represent the topics of the paragraphs. By considering word order, paragraph vectors also address the weaknesses of bag-of-words models in LDA and pLSA. Therefore, paragraph vectors are considered more accurate representations of the context of the content. We can then input resulting vectors

into the analysis using machine learning and clustering techniques for finding similar articles in different academic subjects as well as the relationships between projects from different agencies. Thus, we tried to convert the natural sentences in project descriptions and article abstracts to paragraph vectors in this study.

3. Paragraph embedding using information entropy

This section introduces our proposed paragraph embedding method using entropy and then evaluates whether the similarity of the resulting vectors accurately represents the content similarity of documents.

3.1. Proposal of the paragraph embedding method

Before introducing the proposed method, we present a problem in applying the paragraph vectors for research project descriptions. We implemented the paragraph embedding technique using the Deep Learning Library for Java (https://deeplearning4j.org). Then, we constructed paragraph vectors for approx. 30,000 NSF projects mentioned in the next section. Although we need a more systematic way, but this time the hyperparameters were set empirically as follows: 500 dimensions were established for 66,830 words that appeared more than 5 times; the window size c was 10, and the learning rate and minimum learning rate were 0.025 –0.0001, respectively, with an adaptive gradient algorithm. The learning model is a distributed memory model with hierarchical softmax.

However, the result showed that projects are scattered and not clustered by any subject or discipline in the vector space. Most projects are slightly connected to a low number of projects. Thus, it is difficult to grasp trends and compare an ordinary classification system. Closely observing the vector space reveals some of the reasons for this *unclustered* problem: each word with nearly the same meaning has slightly different word vectors, and shared but unimportant words are considered the commonality of paragraphs. In fact, Le and Mikolov reported classification accuracy with multiple categories of less than 50% [13].

Therefore, for addressing this problem, we introduce the information entropy [14] for clustering word vectors before constructing paragraph vectors. The fact that synonyms tend to gather in a word vector space indicates that the semantics of a word spatially spread to a certain distance. This observation is also suggested in the related literature [15]. Therefore, to unify word vectors of almost the same meanings, excluding trivial common words, we generated clusters of the word vectors based on the semantic diversity of each concept in a thesaurus. We first extracts 19,685 hypernyms (broader terms) with one or more hyponym (narrower term) from the Japan Science and Technology Agency (JST) science and technology thesaurus [16]. The JST thesaurus primarily consists of keywords that have been frequently indexed in 36 million articles accumulated by the JST since 1975. Currently, this thesaurus is updated every year and includes 276,179 terms with English and Japanese notations in 14 categories from bioscience to computer science and civil engineering. Based on the World Wide Web Consortium (W3C) Simple Knowledge Organization System (SKOS), the JST thesaurus

also exists in W3C Resource Description Framework (RDF, https://www.w3.org/RDF/) format with semantic relationships SKOS: broader, SKOS: narrower, and SKOS: related. A broader or narrower relationship essentially represents an *is-a* subsumption relationship but sometimes denotes a *part-of* relationship in geography, body organ terminology, and other academic disciplines. The JST thesaurus is publicly accessible from Web APIs on the J-GLOBAL website (http://jglobal.jst.go.jp/en/), along with the visualization tool Thesaurus Map (http://thesaurus-map.jst.go.jp/jisho/fullIF/index.html). We then calculate the information entropy of each concept in the JST thesaurus from the dataset. Shannon's entropy in information theory is an estimate of event informativeness. We used this entropy to measure the semantic diversity of a concept [17]. After creating clusters according to the degree of entropy, we unify all word vectors in the same cluster to a cluster vector and constructed paragraph vectors based on the cluster vectors. The overall flow is shown in **Figure 1**.

Hereafter, the "word" is a word in the dataset, the "term" is a term in a thesaurus, and terms are classified into hypernyms, hyponyms, and their synonyms. The "concept" is defined as a combination of a hypernym and one or more hyponyms one level below the hypernym indicated as a red box in **Figure 2**. Given that a thesaurus consists of terms T_i, we calculated the entropy of a concept C by considering the appearance frequencies of a hypernym T_0 and its hyponyms $T_1 \ldots T_n$ as an event probability. The frequencies of synonyms $S_{i0} \ldots S_{im}$ of term T_i was summarized to a corresponding concept (synonyms S_{ij} include descriptors of terms T_i themselves).

$$H(C) = -\sum_{i=0}^{n}\left(\sum_{j=0}^{m} p\left(S_{ij} \mid C\right) \cdot \log_2 \sum_{j=0}^{m} p\left(S_{ij} \mid C\right)\right) \tag{3}$$

In Eq. (3), $p(S_{ij} \mid C)$ is the probability of a synonym S_{ij} given a concept and terms T_i. For each concept in the thesaurus, we calculated the entropy $H(C)$ in the dataset. As the probabilities of events become equal, $H(C)$ increases. If only particular events occur, $H(C)$ is reduced because of low informativeness. Thus, the proposed entropy of a concept increases when a hypernym

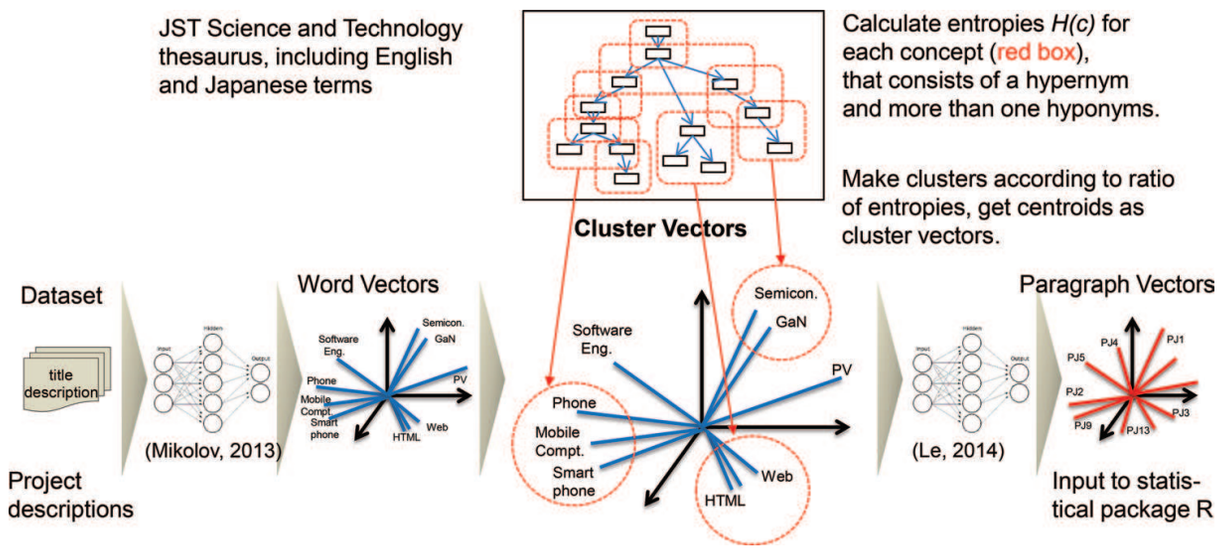

Figure 1. Construction of paragraph vectors based on cluster vectors.

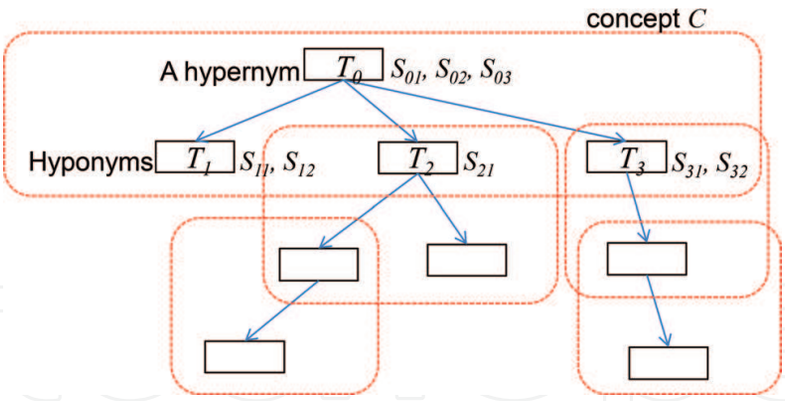

Figure 2. Concepts in a thesaurus.

and hyponyms that construct a concept separately appear with a certain frequency in the dataset. Therefore, the degree of entropy indicates the semantic diversity of a concept. Then, assuming that the degree of entropy and the spatial size of a concept in a word vector space are proportional to a certain extent, we split the word vector space into clusters. In fact, our preliminary experiment indicated that the entropy of a concept has high correlation $R = 0.602$ with the maximum Euclidean distance of hyponyms in the concept in a vector space, at least while the entropy is rather high. Specifically, we refined clusters by repeatedly subdividing them until the defined criterion was satisfied. In our method, we set the determination condition as shown in Eq. (4).

$$Cl(w_k) = \begin{cases} Cl(w_i) & \left(\dfrac{H(C(w_i))}{H(C(w_j))} > \dfrac{\| w_k - w_i \|}{\| w_k - w_j \|} \right) \\ Cl(w_j) & \text{(otherwise)} \end{cases} \tag{4}$$

This condition represents that the word vectors $w_0 \ldots w_T$ are subdivided into two clusters proportionally to the ratio of the highest two concept entropies $H(C(w_i))$ and $H(C(w_j))$, which are selected from all entropies of concepts in a cluster (an initial cluster is the whole vector space). $C(w_i)$ and $C(w_j)$ mean concepts C to which words w_i and w_j belong, respectively. The words w_i and w_j are words, whose lemmatized forms are identical to terms or synonyms in the thesaurus. However, note that the entropies of the other words whose correspondences are not included in the thesaurus are not calculated in Eq. (3). $Cl(w)$ means a cluster to which a vector of a word w should be classified.

The vector space is subdivided until the entropy becomes lower than 0.25 (the top 1.5% of entropies) or the number of elements in a cluster is lower than 10. These parameters were also determined empirically through the experiments. After generating 1260 clusters from 66,830-word vectors, we considered the centroid of all vectors in a cluster as a cluster vector. Then, we constructed paragraph vectors using the cluster vectors rather than word vectors, as shown in Eq. (5) that is an extension of Eq. (2). After all, each cluster vector represents a concept that has the highest entropy in all concepts included in the cluster.

$$L = \sum_{t=1}^{T} \log p\left(Cl(w_t) \mid Cl(w_{t-c}), \ldots, Cl(w_{t+c}), d_i \right) \tag{5}$$

3.2. Evaluation of paragraph vectors

Next, we evaluate the resulting vectors on the map constructed from the following dataset. In this article, the dataset includes titles and abstracts of 266,772 IEEE conference articles published from 2012 to 2016, including 2,290,743 sentences in total and titles and descriptions of 34,192 NSF projects from 2012 to 2016, including 730,563 sentences in total. Note that IEEE journal, transaction, symposium, and workshop articles are not included, and NSF project domains are limited to Computer and Information Science and Engineering, Mathematical and Physical Sciences, and Engineering in accordance with IEEE articles. All words in the sentences were tokenized and lemmatized by Stanford CoreNLP before creating the vector space.

In terms of the *unclustered* problem, we confirmed that the proposed method successfully formed several clusters compared with the original paragraph embedding method. For a quantitative comparison, in **Figure 3** shows the relationships between the cosine similarities and the number of edges, and the relationship between the degree centrality and the number of nodes (i.e., projects) in the case of the cosine similarities of >0.35. As a result, we confirmed that edges with a higher cosine similarity and nodes with higher degrees increase. The reason for this result is because, through the use of high-entropy concepts, which are significant in scientific and technological contexts excluding scientifically unimportant words—as elements between paragraph vectors, the paragraph vectors were able to comprise meaningful groups. Simultaneously, newly, unknown synonyms, and closely related words that are not defined in the thesaurus can be unified to a cluster vector, if they are in the same cluster. Taking the centroid vector as a representative vector in a cluster involves separating each cluster vector as much as possible to form a clear difference in the vector space.

In terms of the accuracy of content similarities, the evaluation encounters difficulty since, to the best of our knowledge, there is no gold standard for evaluating the similarity among scientific and technological documents. Therefore, we first evaluated the degree of the similarities based on a sampling method. We randomly extracted 100 pairs of projects with a cosine similarity of >0.5 (similarities less than 0.5 are not considered in the map layout), to make the distribution similar to the entire distribution. Each pair has two project titles and descriptions, and a cosine value that is divided into three levels: weak ($0.5 \leq \cos. < 0.67$), middle ($0.67 \leq \cos. < 0.84$), and strong ($0.84 \leq \cos.$). Some examples of two projects and their cosine value are shown in **Table 1**. Then, three members of our organization, a funding agency in Japan, evaluated the similarity of each pair. The members were provided the prior explanations for the intended use of the map

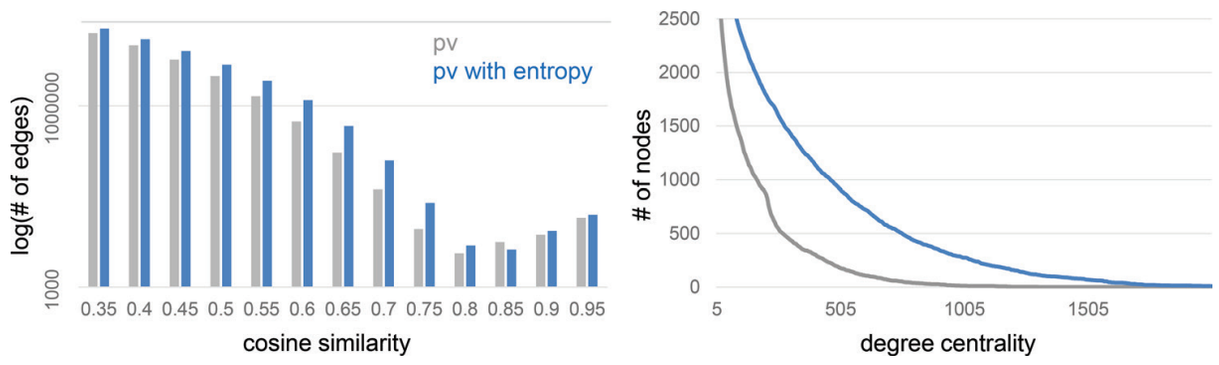

Figure 3. Comparison between paragraph vectors and those with entropy clustering.

title / (desc.)	*cos.*	*title / (desc.)*
understanding the physics of galaxy formation and evolution at high redshift / understanding the processes regulating galaxy…	0.50 (weak)	the birth of the first stars and galaxies / the aim of this proposal is to simulate the formation and evolution of galaxies within the…
asymptotic graph properties / many parts of graph theory have witnessed a huge growth over the last years, partly because of their relation to theoretical computer science and statistical physics. …	0.52 (weak)	benjamini-schramm approximation of groups and graphings / large graphs have become central objects in many fields in the last couple of decades: in neural sciences, network sciences…
a high intensity neutrino oscillation facility in Europe / the recent discovery that the neutrino changes type (or flavour) as it travels through space, a phenomenon referred to as neutrino oscillations,…	0.53 (weak)	probing fundamental properties of the neutrino at the sno+ experiment / i propose a comprehensive programme of research on sno+, a multi-purpose neutrino experiment that has the capacity…
systems biology of pseudomonas aeruginosa in biofilms / systems biology is a new and rapidly growing discipline . it is widely…	0.54 (weak)	cyclic-di-gmp: new concepts in second messenger signaling and bacterial biofilm formation / biofilms represent a multicellular…
investigation of human nucleoporins stoichiometry and intracellular distribution by quantitative mass spectrometry / the nuclear pore complex (npc) is one of the most intricate multi-protein…	0.56 (weak)	atlas of cell-type specific nuclear pore complex structures / the nuclear pore complex (npc) is one of the most intricate components of eukaryotic cells and is assembled from ~30 nucleoporins…
european science and technology in action building links with industry, schools and home / the aim of establish is to facilitate and implement an inquiry based approach in the teaching and learning…	0.67 (middle)	science teacher education advanced methods / helping teachers raise the quality of science teaching and its educational environment has the potential to increase student engagement,…
support to tenth european conference on turbomachinery - fluid dynamics and thermodynamics, lappeenranta, finland, 15-19 march 2013 / the european turbomachinery conference is…	0.99 (strong)	support to ninth european conference on turbomachinery - fluid dynamics and thermodynamics, istanbul, turkey, 21-25 march 2011 / the european turbomachinery conference is…

Table 1. Example of sampled projects/articles.

and some examples of evaluation. The members received the same data, and their backgrounds are bioscience, psychology, and computer science. As a result, we confirmed that 78% of the similarities matched majority votes of the members' opinions. Examples misjudged include, for example, the not related pairs of two projects that have the same acronyms with different meanings, and the stronger pairs of two projects that have only a few common words, but which are recent technologies attracting attention. We expect that those words will eventually have higher entropies and then the project similarities will be estimated to be stronger. We also plan to replace acronyms in project descriptions with full words before making vectors. By contrast, the accuracy of the similarities of the original paragraph embedding method was 21%. The evaluation results were determined to be in "fair" agreement (Fleiss' Kappa $\kappa = 0.29$) (**Table 2**).

Moreover, we evaluated the accuracy of content similarities using the artificial data, part of which is randomly replaced with the other projects/articles. We replaced 10, 20, …, 100% of

Similarity	**Weak**	**Middle**	**Strong**
Precision	77.5	83.3	100.0
Recall	98.6	33.3	83.3
F1 value	86.8	47.6	90.9

Table 2. Evaluation of similarity based on sampling (%).

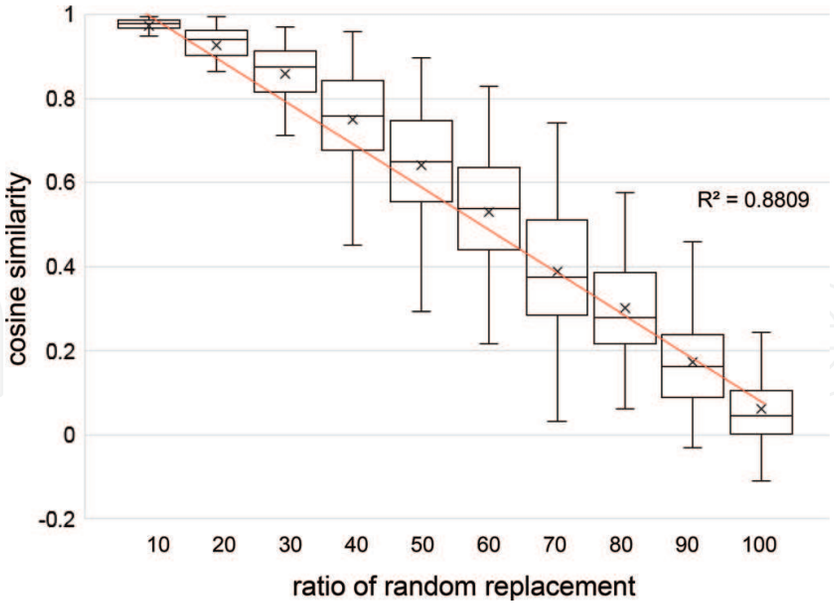

Figure 4. Cosine similarities of artificial data with partial replacement.

a project description or a article abstract with sentences randomly selected from the others. Then, we measured a cosine similarity between a vector generated from the artificial project/article and a vector of the original project/article. The projects/articles were randomly selected from all projects/articles, and then we evaluated 1000 pairs of the original project/article and the artificial project/article. The relationship of the replacement ratios and the cosine similarities is shown in **Figure 4**. As a result, we confirmed that there is an obvious correlation between content similarities of projects/articles and their cosine similarities with $R^2 = 0.89$. The paragraph vectors without the entropy clustering also had the same trend, but the vectors with the entropy clustering had higher similarities on average. This result matches the relationships between the cosine similarities and the number of edges shown in **Figure 3**.

4. Mapping Science

This section describes our content-based map of science, Mapping Science [18, 19]. After introducing its interface, we describe our clustering and layout method of articles and projects in the map and analytical functions provided.

4.1. Interfaces

In **Figure 5** shows three main views of the Mapping Science, which are a portfolio view, a clustered view, and analytic views.

In the portfolio view, five research areas, Information, Mathematics and Physics, Communication, Electronics and Mechatronics, and Power and Energy, to which the entire dataset has been divided by full-text search with predefined queries, are shown. The size of circles corresponds to the number of articles and projects in the area.

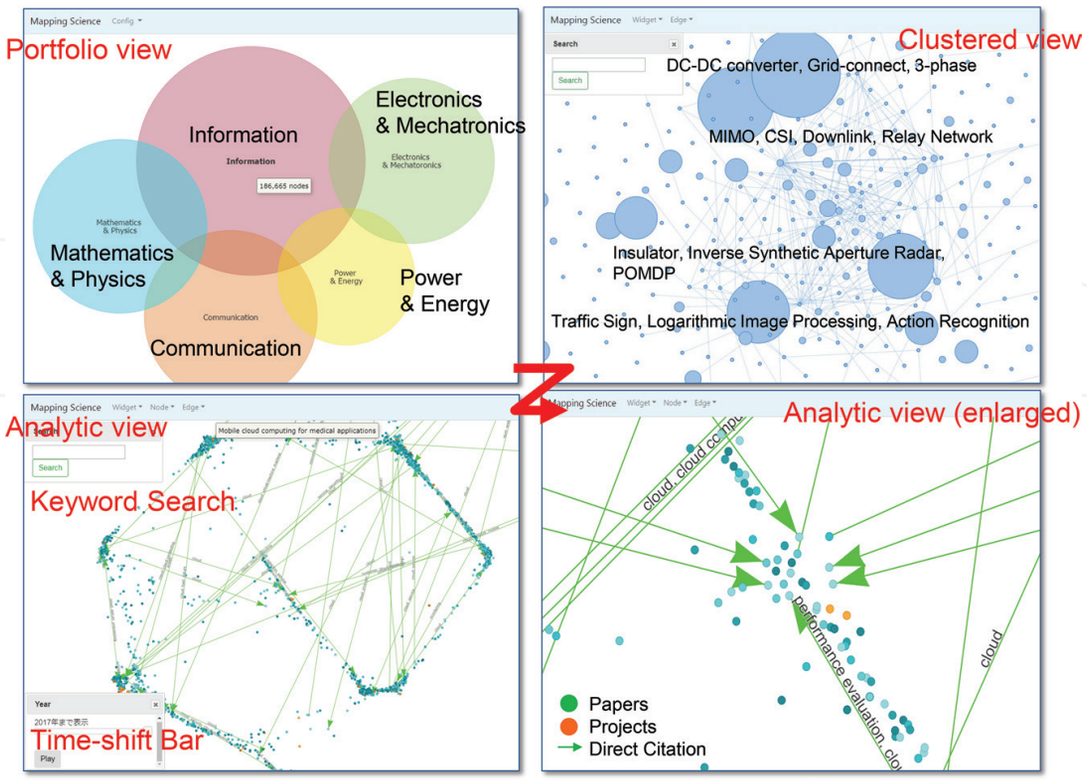

Figure 5. Interface of Mapping Science.

In the clustered view that opens when users click one of the areas in the portfolio view, the results of clustering all the articles and projects in the area are shown. The details of the clustering method are shown in the next section. This view is for taking a look at the technologies in the area. Each cluster has at most 10 labels, which are extracted as feature phrases using a probabilistic information retrieval method, BM25 [20].

In the analytic view that opens when users click one of the clusters in the clustered view, each node corresponds to a article or a project, and distances between the nodes are proportional to the cosine similarities between articles/projects, as much as possible. In addition, direct citation links between articles (citing → cited) are shown in light green edges with labels showing common phrases between two articles, which are also extracted by the BM25 method. When users click a node, the detailed information about the node (article or project) is shown on the map.

In all the views, the search box located at the upper-left corner provides full-text search for all articles and projects included in the current view, and the search results are highlighted in the view. Moreover, the analytic view provides the time-shift bar, which displays the cumulative changes in a cluster according to published/started years of articles/projects. The trial version of this map is publicly available at https://jipsti.jst.go.jp/foresight/.

4.2. Clustering and layout method of the nodes

In this section, we describe a method for generating the clustered view and the analytic view. There are too many nodes (articles and projects) even in a research area to explore a specific research topic (over 160,000 nodes included in the Information area in **Table 3**). We thus

	Information	Mathematics and Physics	Communication	Electronics and Mechatronics	Power and Energy
# of nodes	165,823	113,982	99,995	88,023	89,845
# of clusters	474	345	338	400	303
# of clusters (only by infomap)	2313	1614	1630	2807	1776

Table 3. # of nodes and clusters in each research area.

divided them into several hundred clusters and provided analytic functions described in the next section to explore articles and projects in each cluster.

A major concern in clustering and laying out the nodes is to reduce 500-dimensional paragraph vectors to a 2D network structure. In general, conventional clustering or dimension reduction techniques such as multi-dimensional scaling (MDS) have $O(n^3)$ computational complexity, which increases the calculation time in proportion to that. We thus, to accommodate the practical calculation time, generated a network structure only from the edges that are the 30 highest similarities (at least, 0.5 or more) to other nodes. Sci2Tool [3] also generated the network only from the 15 highest similarities edges and successfully created an informative map of journals.

Clusters in the clustered view are calculated by info map [21], which is one of modularity-based network clustering algorithms [22]. By increasing the modularity, the nodes are divided into clusters that have more edges within the clusters than edges between the clusters. Thus, articles or projects in a cluster have relatively high similarities and form meaningful sets. However, the simple application of the info map generated too many clusters to explore the clustered view (over 2800 clusters included in Electronics & Mechatronics area in **Table 3**). Therefore, we merged small clusters comprised of less than 50 nodes into the nearest cluster, which has the highest similarity pair between any of two nodes in the clusters. This operation corresponds to a single linkage clustering in agglomerative clustering. As a result, the numbers of clusters are reduced as in **Table 3**. Although the accuracy of the clustering result falls (the modularity decreases), nodes incorporated into the nearest cluster tend to form independent sets of nodes in the analytic view and can be distinguished in the view. The distances between clusters in the clustered view mean the distances in the single linkage-clustering.

The layout algorithm in the analytic view is OpenOrd (formally, DrL) [23]. This is a well-known force-directed layout algorithm and frequently used in other maps of science such as Sci2Tool. In **Figure 6** shows a comparison of layout algorithms for Internet of thing cluster (see the next section), which includes the OpenOrd (edge cut parameter: 0.88, 0.91, and 0.94), MDS with cosine dissimilarity, large graph layout (LGL) [24] and Fruchterman Reingold layout (FR) [25]. The LGL and the FR are also force-directed algorithms. We can obviously confirm several clusters in the OpenOrd, but those are not clear in the other algorithms. The number of clusters in the OpenOrd increase as the edge cut parameter increases. Thus, we empirically set the OpenOrd with the edge cut parameter: 0.91 in the analytic view by default. The other parameters were also empirically set to show the structural features as much as possible. However, as shown in the next section, the analytic view provides several other layout algorithms and parameters; thus, users can change the layout of nodes according to their needs.

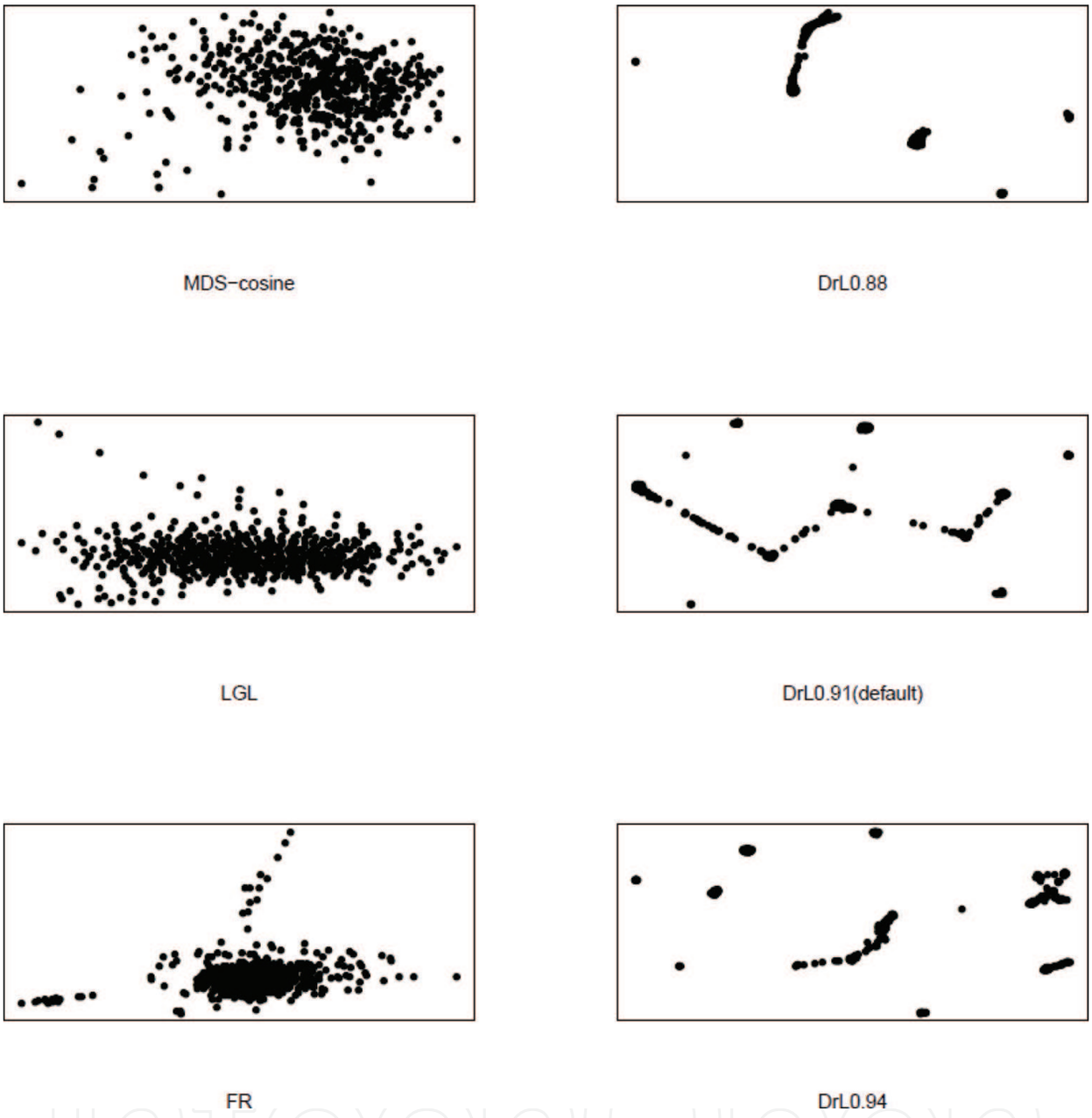

Figure 6. Comparison of graph layout algorithms.

4.3. Analytical functions provided on the map

In addition to the functions described in Section 4.1, the Mapping Science provides the following analytical functions: (1) translation of article abstracts and project descriptions, (2) visualization of statistical information, (3) summarization of feature phrases, (4) querying and exporting using SPARQL, (5) change of layout algorithms, and (6) generation of customized analytic views.

4.3.1. Abstract/description translation function

In the analytic views, users can see the detailed information, such as titles, article abstracts/ project descriptions, authors/project members, affiliations, and publication year/proposed

year. In addition, the abstracts/descriptions are translated into Japanese by clicking "Translate" buttons. The users can read the original abstracts/descriptions in the same pane for confirming the translation validity.

4.3.2. Visualization function of statistical information

As in **Figure 7**, the analytic view can visualize the summary of bibliometric information of the nodes contained in the view. There are several widgets, such as for citation (Impact Factor, SJR, and CiteScore) metrics, publications by year, citations by year, and publications by each country. Moreover, the users can select the nodes in a rectangle area and see the statistical information of the selected nodes. The upper part of the publication by country shows an article count (AC) (https://www.natureindex.com/faq). The AC means the country-level participation in a study, where a country is counted if one or more authors of the article are from the country. For example, if countries of three authors' affiliations in a article are A, B, and B, A is counted as one and B is also counted as one. In contrast, the lower part of the publication by country shows a fractional count (FC) that means the contribution of each country. In the above example, A becomes 1/3, B becomes 2/3.

4.3.3. Summarization function of feature phrases

As in **Figure 8**, the feature phrases of the selected nodes can be summarized in word clouds. At most 10 feature phrases of each node are extracted based on the BM25 method in advance. Then, if the users select the multiple nodes, the feature phrases with higher frequencies are displayed larger and placed closer to the center of the word cloud. This function is useful for understanding specific themes of the selected nodes in a cluster.

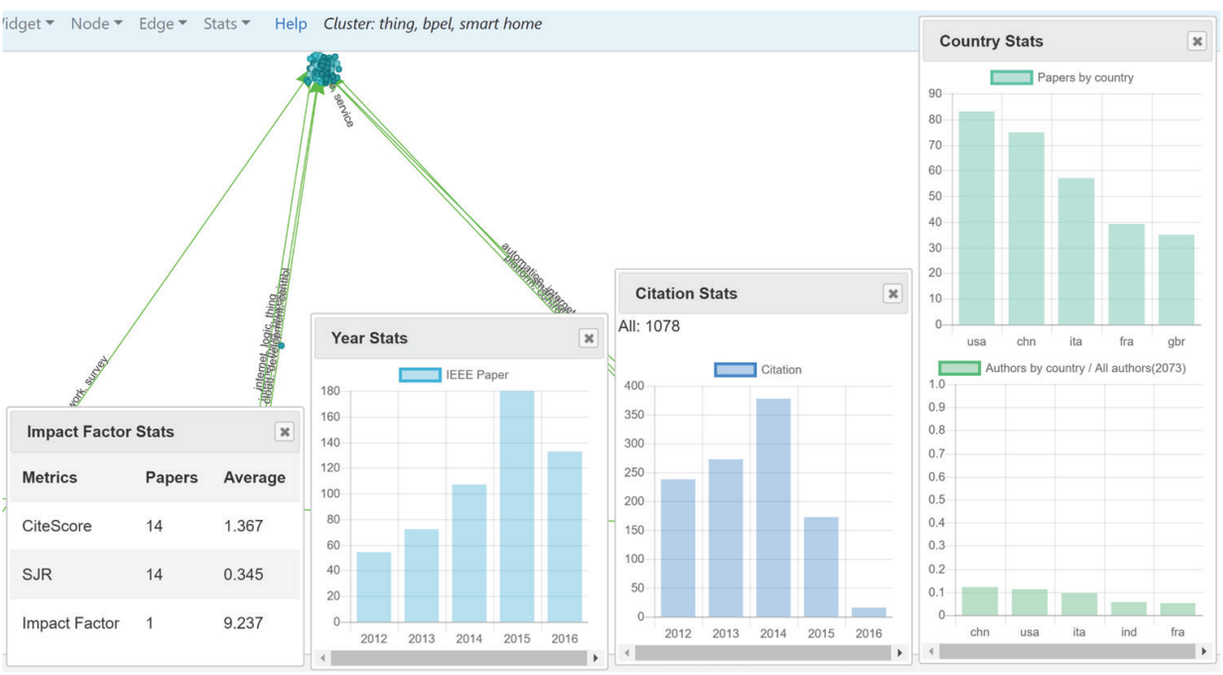

Figure 7. Statistical information.

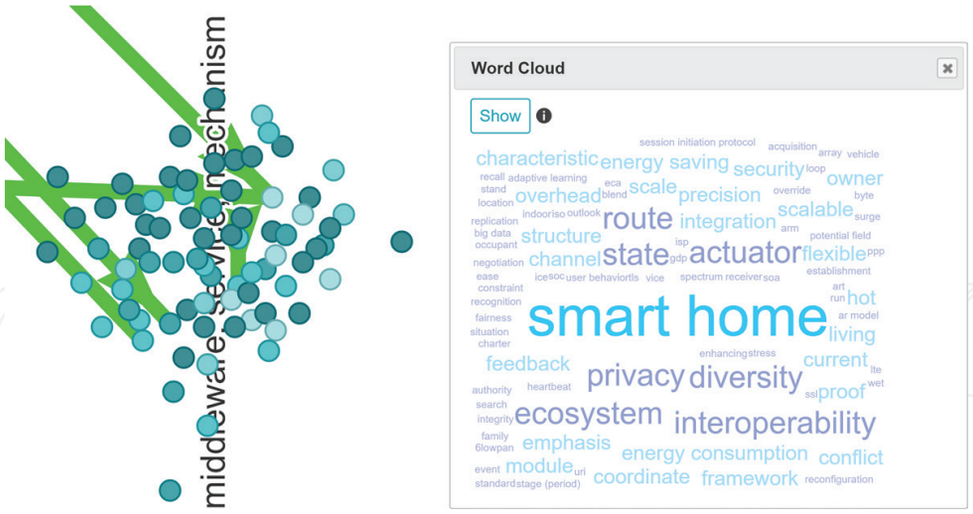

Figure 8. Feature phrases in the selected nodes.

4.3.4. Query function and export function

The background data in the Mapping Science have been converted to RDF data and stored in a graph database. Therefore, the analytic views provide a high-level search using a formal query language, SPARQL, as in **Figure 9**. For example, the users can search for articles, which have >0.8 similarities with articles cited 100+ times from journals with >10 impact factor (such articles might be obscure but important). When the users click a node ID in the result table,

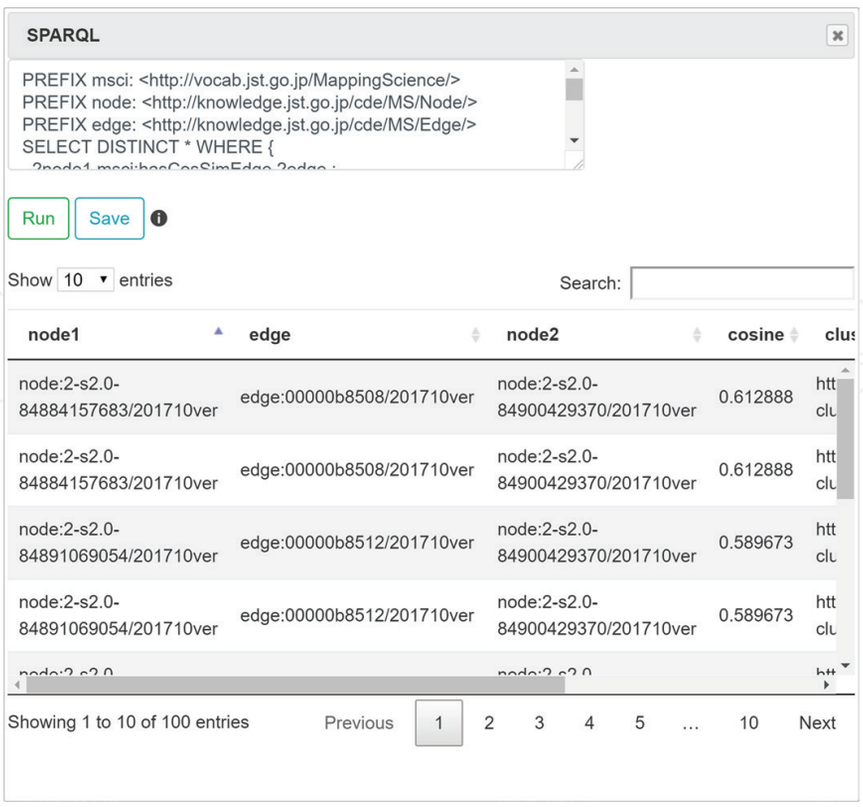

Figure 9. SPARQL search widget.

the node is highlighted and the viewpoint is automatically moved to the node. Moreover, the users can store their own SPARQL queries as macros. Therefore, users who are not familiar with SPARQL can simply call the macros and obtain the query results.

In addition, since we received requests for downloading the information displayed on the map, the information of the selected nodes and all nodes in a cluster can be exported in comma-separated values (CSV) format. The result of SPARQL queries can be also exported in CSV format.

4.3.5. Layout change function

As described in the previous section, the layout of the analytic view was calculated by the OpenOrd (edge-cutting value: 0.91). In addition to that, the analytic views can be redrawn by the OpenOrd (edge-cutting value: 0.94 or 0.88), LGL, Fruchterman-Raingold, or Kamada-Kawai [26]. When the users select a layout, the layout algorithm is executed in the background, the resulting layout information is stored and the view is redrawn. If the layout information is stored in advance, the layout is redrawn immediately. The layout calculation time depends on the number of nodes, and the average time is a few seconds to a few minutes.

4.3.6. Custom analytic view function

The analytic views were composed by the info map algorithm, but the users can create the customized Analytic views by keyword search. When the users enter keywords into the widget in the portfolio view, the nodes are extracted by the full-text search for all nodes in five research areas, and then the layout is calculated by the OpenOrd based on the cosine similarities of the extracted nodes. For example, an analytic view for studies related to neural networks and artificial intelligence across multiple research areas can be created by keywords such as "Artificial Intelligence [AND] Neural Network." This function could help find interdisciplinary studies. The calculation time depends on the number of nodes, and the average time is a few seconds to a few minutes. The information on the customized analytic views is stored in the background; the same view is immediately displayed for the second time. The customized analytic view can provide the same analytical functions, such as keyword search, visualization of statistical information, visualization of the cumulative changes by year, and layout change.

5. Case study for the formation process of research areas

In this map, we try to understand the formation processes of several research areas through chronological changes of network structure. This section describes two cases for the Internet of Things (IoT) and Brain-Computer Interface (BCI).

In **Figure 10** shows the analytic views for an IoT area from 2012 to 2016, which includes 574 nodes as of 2016. The last view is the analytic view in 2016 displaying >0.6 cosine similarities as edges.

In 2012, four islands (places, at which nodes are densely located) mainly for IoT frameworks and networks and for IoT system and security are barely found (labels of each island have been extracted by the summarization function of feature phrases).

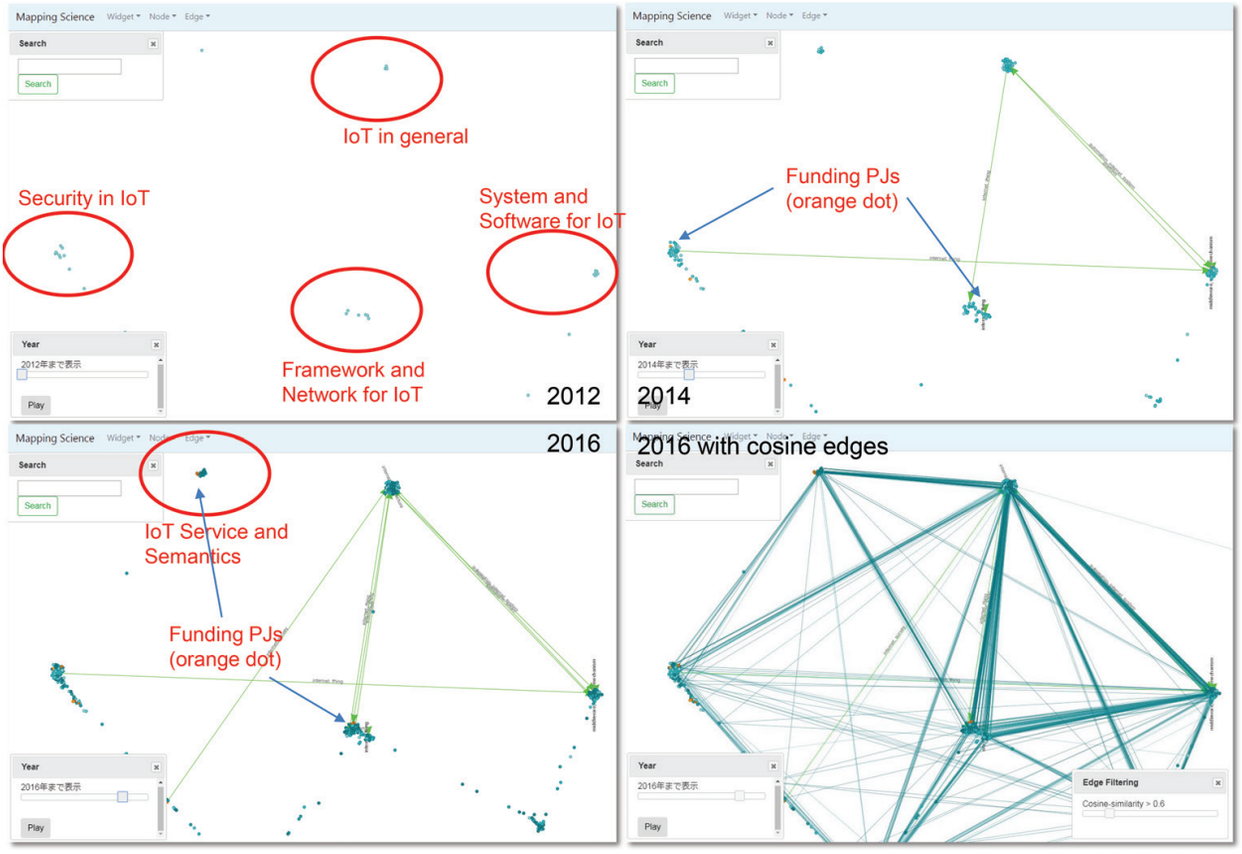

Figure 10. Formation of IoT areas.

In 2013, a funding project (orange node) was firstly established in the security, and then the corresponding island grew bigger, that is, the number of articles increased, although a causal relationship is unclear.

Then, in 2014, the island of the IoT frameworks and networks also had a funding project and grew bigger. At the same time, researchers of each island, which seem to correspond to the different research community, started to recognize with each other, and thus mutual citation links (light green edges) between islands began to be drawn.

In 2015 and 2016, this movement was accelerated; thus, we can confirm that the islands were getting bigger and denser, and mutual citation links increased. Moreover, the other islands than the first four islands, for example, an island for IoT services and semantics at the upper-left corner also gradually grew, and some of them are greatly increasing the articles by getting funding projects.

Finally, the edges of the cosine similarity 0.6 in the last view mean relatively weak similarity described in Section 3.2. In contrast, nodes which compose an island are mutually connected with stronger similarities, although they are too dense to confirm in the figure. Therefore, in this IoT area, there are several research communities dedicated to specific research themes, and they are mutually connected with their content similarity and citation relations. Thus, we can understand that they are developing each theme while forming the IoT area as a whole.

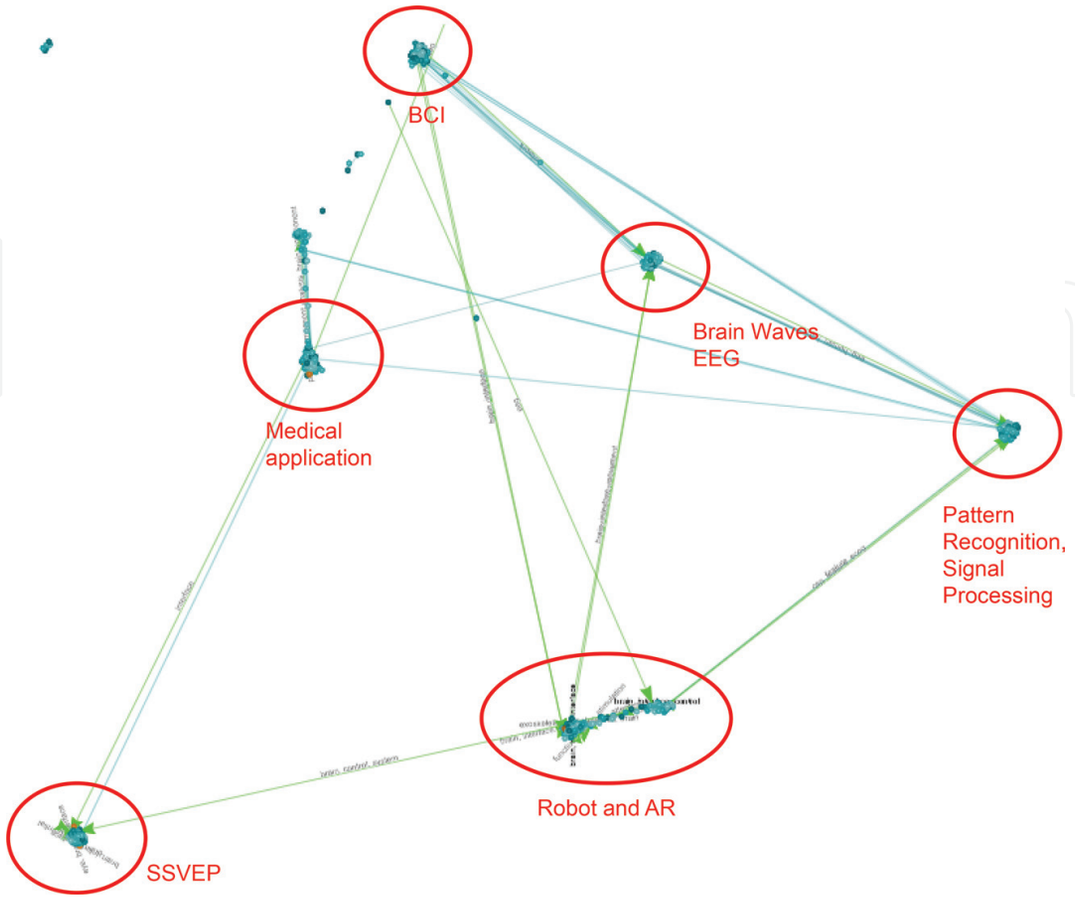

Figure 11. Formation of BCI.

We confirmed several other processes of research area formation in our case studies. For example, in **Figure 11** shows the analytic view for BCI in 2016. In this figure, an island at the top is growing while citing articles for several specific research themes, such as medical applications, brain waves, pattern recognition, and steady state visual evoked potentials (SSVEP). Thus, we can understand that the BCI has been simultaneously approached from several different conventional research themes, and is integrating them. In this manner, we confirmed that the formation processes of research areas can be captured by closely observing the map.

6. Conclusion and future work

In this study, we developed a map of science, Mapping Science based on the research content similarity for funding project descriptions and recently published articles, which have difficulty in applying the citation analysis. After improving the existing paragraph embedding technique with an entropy-based clustering method of word vectors, we confirmed the good face validity. Then, we introduced the map constructed from approx. 300 k IEEE articles and NSF projects from 2012 to 2016 with the clustering and layout method of articles/projects and analytic functions provided on the map. Finally, we confirmed that formation processes of some specific research areas can be captured as changes of network structure.

As the next step, we plan to have a comparison with citation-based methods on concrete scenarios and incorporate patent information on the map. In addition, by overlaying domestic funding projects with NSF and Horizon2020 through the JST thesaurus that has English and Japanese notations, we will identify the trend of public grants. Finally, we try to extract metrics from chronological changes of the network structure of research areas. Foresight and understand from scientific exposition (FUSE) program in Intelligence advanced research projects activity (IAPRA) already conducted a study for identifying emerging research area based on several metrics obtained from several maps of science from 2011 to 2015. We, JST, will also utilize such metrics in statistical analysis and machine learning techniques to detect emerging research areas in their early stage for the next science and technology policies.

Author details

Takahiro Kawamura*, Katsutaro Watanabe, Naoya Matsumoto and Shusaku Egami

*Address all correspondence to: takahiro.kawamura@jst.go.jp

Japan Science and Technology Agency, Tokyo, Japan

References

[1] Price D. Networks of scientific articles. Science. 1965;**149**:510-515

[2] Borner K. Sci2: A tool of science of science research and practice. In: Tutorial of the 10th International Conference on Scientometrics and Informetrics (ISSI 2011); 2011

[3] Boyack K, Klavans R, Borner K. Mapping the backbone of science. Scientometrics. 2005; **64**(3):351-374

[4] Steyvers M, Griffiths T. Probabilistic topic models. In: Landauer T, McNamara D, Dennis S, Kintsch W, editor. Latent Semantic Analysis: A Road to Meaning. Hillsdale, NJ: Laurence Erlbaum; 2007

[5] Blei D, Ng A, Jordan M. Latent dirichlet allocation. Journal of Machine Learning Research. 2003;**3**:993-1022

[6] Griffiths T, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences. 2004;**101**(suppl. 1):5228-5235

[7] Talley E, Newman D, Mimno D, Herr B II, Wallach H, Burns G, Leenders A, McCallum A. Database of NIH grants using machine-learned categories and graphical clustering. Nature Methods. 2011;**8**:443-444

[8] Herr II B, Talley E, Burns G, Newman D, LaRowe G. The NIH visual browser: An interactive visualization of biomedical research. In: Proceedings of 13th International Conference on Information Visualisation (ICIV 2009); 2009. pp. 505-509

[9] Kullback S, Leibler R. On information and sufficiency. Annals of Mathematical Statistics. 1951;**22**:79-86

[10] Firth JR. A synopsis of linguistic theory 1930-1955. Studies in Linguistic Analysis. 1957;**1952-59**:1-32

[11] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proceedings of Workshop at the International Conference on Learning Representations (ICLR 2013); 2013

[12] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS 13). Vol. 2; 2013. pp. 3111-3119

[13] Le Q, Mikolov T. Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014). 2014;**32**(2):1188-1196

[14] Shannon C. A mathematical theory of communication. Bell System Technical Journal. 1948;**27**(379-423):623-656

[15] Vilnis L, McCallum A. Word representations via Gaussian embedding. In: Proceedings of International Conference on Learning Representations (ICLR 2015); 2015. pp. 1-12

[16] Kimura T, Kawamura T, Watanabe K, Matsumoto N, Sato T, Kushida T, Matsumura K. J-GLOBAL knowledge: Japan's largest linked data for science and technology. In: Proceedings of the 14th International Semantic Web Conference (ISWC 2015); 2015

[17] Santus E, Lenci A, Lu Q, Walde S. Chasing hypernyms in vector spaces with entropy. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014); 2014. pp. 38-42

[18] Kawamura T, Watanabe K, Matsumoto N, Egami S, Jibu M. Funding map for research project relationships using paragraph vectors. In: Proceedings of the 16th International Conference on Scientometrics & Informetrics (ISSI 2017); 2017. pp. 1121-1131

[19] Kawamura T, Watanabe K, Matsumoto N, Egami S, Jibu M. Science graph for characterizing the recent scientific landscape using paragraph vectors. In: Proceedings of the 9th ACM International Conference on Knowledge Capture (K-Cap 2017); 2017. pp. 9-16

[20] Jones KS, Walker S, Robertson SE. A probabilistic model of information retrieval: Development and comparative experiments. Information Processing and Management. 2000;**36**(6):779-808

[21] Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences of the United States of America (PNAS 2008). 2008;**105**(4):1118-01123

[22] Newman MEJ. Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America (PNAS 2006). 2006; **103**(23):8577-8582

[23] Martin S, Brown WM, Klavans R, Boyack K. OpenOrd: An open-source toolbox for large graph layout. In: Proceedings of SPIE, Visualization and Data Analysis (VDA); 2011. p. 786806

[24] Adai AT, Date SV, Wieland S, Marcotte EM. LGL: Creating a map of protein function with an algorithm for visualizing very large biological networks. Journal of Molecular Biology. 2004;**340**(1):179-190

[25] Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. Software—Practice and Experience. 1991;**21**(11):1129-1164

[26] Kamada T, Kawai S. An algorithm for drawing general undirected graphs. Information Processing Letters. 1989;**31**(1):7-15