frontiers in Big Data

ORIGINAL RESEARCH
published: 09 January 2020
doi: 10.3389/fdata.2019.00048

Application of a Novel Subject Classification Scheme for a Bibliographic Database Using a Data-Driven Correspondence

Kei Kurakawa[1]*, *Yuan Sun*[2] *and Satoko Ando*[3]

[1] *Scholarly and Academic Information Division, Cyber Science Infrastructure Development Department, National Institute of Informatics, Tokyo, Japan,* [2] *Information and Society Research Division, National Institute of Informatics, Tokyo, Japan,* [3] *Clarivate Analytics (Japan), Co. Ltd., Tokyo, Japan*

A novel subject classification scheme should often be applied to a preclassified bibliographic database for the research evaluation task. Generally, adopting a new subject classification scheme is labor intensive and time consuming, and an effective and efficient approach is necessary. Hence, we propose an approach to apply a new subject classification scheme for a subject-classified database using a data-driven correspondence between the new and present ones. In this paper, we define a subject classification model of the bibliographic database comprising a topological space. Then, we show our approach based on this model, wherein forming a compact topological space is required for a novel subject classification scheme. To form the space, a correspondence between two subject classification schemes using a research project database is utilized as data. As a case study, we applied our approach to a practical example. It is a tool used as world proprietary benchmarking for research evaluation based on a citation database. We tried to add a novel subject classification of a research project database.

Keywords: bibliographic database, data-driven correspondence, research project database, subject classification scheme, topological space

OPEN ACCESS

Edited by:
Feng Xia,
Dalian University of Technology (DUT), China

Reviewed by:
Michele A. Brandão,
Federal Institute of Minas Gerais, Brazil
Silvio Simani,
University of Ferrara, Italy

***Correspondence:**
Kei Kurakawa
kurakawa@nii.ac.jp

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *31 August 2019*
Accepted: *29 November 2019*
Published: *09 January 2020*

Citation:
Kurakawa K, Sun Y and Ando S (2020) Application of a Novel Subject Classification Scheme for a Bibliographic Database Using a Data-Driven Correspondence. Front. Big Data 2:48. doi: 10.3389/fdata.2019.00048

INTRODUCTION

Subject classification is a popular and useful aspect for academic database and data analysis. Academic resources, such as research articles, journals, conference proceedings, books, field samples, software, and various electronic materials, are organized by subject classifications in the general or domain-specific approach. University libraries, institutional resource centers, and research labs organize their research resources in an efficient manner to gain easy access to these resources when necessary. Research funding agencies manage their applicants, projects, and reports by classifying research subjects, which are often diversified and transformed to reflect on the current research landscape. Academic fields are fundamental concepts of academic classifications for organizing academic materials. From analysis perspectives, institutional research (IR) focuses on research and educational activities, in which the research and educational portfolios of researchers, professors, and staff are analyzed through subject classifications. Moreover, the databases of national grants are often surveyed via subject classifications.

Practically, classification has been utilized in library catalogs for not less than a 100 years (Hjørland, 2008). The Dewey Decimal Classification is an old library classification invented in 1876 and is popular for classifying books in the shelves of university libraries. Other popular library

classifications, such as the Universal Decimal Classification, the Library of Congress Classification, and the Colon Classification, were also invented a 100 years ago. They were revised several times to fit with the current book subject diversity. Japanese library classification examples include the Nippon Decimal Classification and the Japan National Diet Library Classification, which were released in 1928 and 1963, respectively. For academic journals, the Web of Science (WoS) subject classification is one of the most popular subject classifications for the WoS citation database. For research evaluation purposes, journals are frequently being classified based on specific viewpoints. The Essential Science Indicator (ESI) is one of the specifically developed subject classifications for research evaluation based on the WoS citation database.

In the research evaluation domain, special subject classifications that are developed for research and educational work should be adopted throughout all kinds of target databases (Gómez et al., 1996). National research and educational evaluation organizations use their original subject classifications to classify organizations and persons suitable for domestic evaluation tasks. Then, they compare them globally based on research and educational output records collected from world common output databases such as the WoS citation database. For example, the UK government defines Units of Assessment as subject classifications for the Research Assessment Exercise and the Research Excellence Framework. The Italian evaluation agency for university and research systems ANVUR (Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca) constructed an original category scheme to be used for its evaluations. Excellence in Research for Australia, which is an Australian research evaluation program, developed the original subject classification scheme: Fields of Research. All these subject classifications must be adopted for the WoS citation database to analyze their national activities and compare them with regard to their common standards. Along with current international business qualifications on research evaluation, the same requirements emerged from the universities in Japan to ensure that the Japanese national funding programs KAKENHI subject classifications would be adopted for the WoS citation database.

However, adopting subject classifications for bibliographic databases is a highly challenging task. For example, in 2019, the WoS citation database in InCites™, which is a research output evaluation tool, comprised 58,395,008 article records of 24,688 journals. Even assigning subject categories for articles or journals as a set of units is labor intensive and time consuming. Hence, an excellent, effective, and efficient method for assigning their subject categories is necessary. In this study, we propose an approach to applying a novel subject classification scheme for the WoS citation database.

Our main contributions of the work are as follows:

- We propose an approach to apply a novel subject classification scheme for a subject-classified database using a data-driven correspondence between the new and present ones, which is accustomed to digital libraries.
- We give a fundamental analytical model of subject classification scheme based on set theory and describe compact topological space formation for a new subject classification scheme as a necessary condition.
- We demonstrate the effectiveness and efficiency of our approach to a practical bibliographic database.

In the following sections, firstly we look around the related work to state our approach in section Related Work, then describe our approach in the section Data-Driven Approach to Applying a Novel Subject Classification Scheme, and successively demonstrate the effectiveness and efficiency of our approach in a case study in section Case Study. Finally, we conclude our approach in the section Conclusions and Future Work.

RELATED WORK

In the above introductory section, we mentioned an issue around subject classifications from the viewpoint of library and information science and scientometrics. To tackle this issue, a series of related work in the computer science such as information retrieval, data mining, and digital libraries have been investigated for several decades. As for the general problem setting, subject classification of research items on bibliographic database refers to a part of automated text categorization problems. It goes back to Maron's (1961) seminal work on a probabilistic text classification. The methods are mainly divided into two types, i.e., supervised learning and unsupervised learning, which are also named as classification and clustering. The former requires the labeled data indicating the right answer to a given decision problem so as to derive classifiers. The classifiers then are applied to the target data to be classified. The example methods are naïve bays classification, neural networks, support vector machines. The latter does not need such the labeled data and extracts intrinsically the classification pattern from data and classify them. The example methods are k-means, expectation maximization (EM), hierarchical agglomerative clustering, divisive clustering, matrix decompositions, e.g., latent semantic indexing (LSI) and principal component analysis (PCA), and topic modeling, e.g., latent Dirichlet allocation (LDA). Sebastiani (Sebastiani, 2002) thoroughly surveyed these classifier techniques from the computer science perspective, and Jain et al. (1999) also surveyed clustering techniques for computer vision. Both classification and clustering are so general that they are frequently organized and explained from different basic contexts, such as pattern recognition (Bishop, 2006), information retrieval (Manning et al., 2008) and data mining (Han et al., 2011).

When adapting a method to the predefined classification scheme, classification is better utilized than clustering. Classification learns a decision from a labeled data, in contrast, clustering learns implicit relationships of unlabeled data. In relation to our problem setting, multi-label classification or multi-label learning have been investigated on several basic machine learning architectures. Multi-label classification classifies target data under $2^{|L|}$ classification space where L is a set of labels. Recent examples are multi-label learning based

on SVM (Chang et al., 2017), based on deep learning (Mai et al., 2018), and based on ensemble classification (Büyükçakir et al., 2018). For very large classification space, extreme multi-label classification is proposed, e.g., a method based on graph embedding (Tagami, 2017), a method based on convolutional neural network (CNN) (Liu et al., 2017), and a method based on attention model of neural networks (Wang et al., 2018). Moreover, label hierarchy also can be considered so that part-of, is-a, and inclusion relationships are extracted from external data sources such as Wikipedia in the classification task (Bairi et al., 2016; Xie et al., 2017).

In digital libraries, the mappings between different classification schemes have been considered for a long period. For example, the method of automatically converting from existing classifications of documents into another scheme used in a quality-controlled database is occasionally used in co-operative cataloging projects and union catalogs, sometimes even in individual OPACs as soon as cataloging records using a different classification scheme are imported or exchanged (Koch et al., 1997). These mappings for the purpose of information integration and exchange was widely discussed in the 1970s, and is even more relevant today with the overall trend of information integration on the web (Slavic, 2011).

DATA-DRIVEN APPROACH TO APPLYING A NOVEL SUBJECT CLASSIFICATION SCHEME

Our approach is accustomed to digital libraries. We assume that a subject classification scheme has been originally adopted for a bibliographic database such as the WoS citation database. Then, we attempt to apply a new subject classification scheme for this database based on the relationship between two subject classification schemes. The relationship is the correspondence between them, which is induced by data.

Subject Classification Model of the Bibliographic Database

First, we defined the subject classification model of the bibliographic database such as the WoS citation database to explain our approach. This model is a mathematical formula and a psychological aspect of subject categories embedded in the database.

Then, we assumed that a bibliographic database representing a set of articles for scientific research is available. Each article is labeled with at least one category of the subject classification scheme. That is, all articles are classified on the basis of the subject classification scheme. This scheme implies compact topological space in the database. It states the database structure that affects the analysis by using the subject classification scheme.

Definition 1 (database with a subject classification scheme). A database S is a set of articles a_n, and a subject classification scheme C is a set of subject categories c_λ. Articles attributed to a subject category comprise a subset of S; hence, subject categories in a subject classification scheme refer to a family of subsets $(O_\lambda)_{\lambda\in\Lambda}$ of S. O is an open set, whereas Λ is an index set. A subset O_λ depends on the corresponding subject category c_λ. Therefore, we define a map f from the subject classification scheme C to the powerset $\mathfrak{P}(S)$.

Theorem 1 (finite cover). A practical subject classification scheme C is mapped to a finite cover $\mathfrak{O}$ of S.

Proof. In practical bibliographic databases, a subject classification scheme C consists of finite elements c_λ that are mapped to finite subsets O_λ using a map f. Let $\mathfrak{O}$ be a subset of $\mathfrak{P}(S)$ comprising $\{O_i|i\in I\}$, where I is a finite index set. Hence, $S=\bigcup_{i\in I}O_i$ ($O_i\in\mathfrak{O}$), where $\mathfrak{O}$ is referred to as a finite cover of S.

Theorem 2 (compact topological space). A practical subject classification scheme C implies a compact topological space $\left(S,\widetilde{\mathfrak{O}}\right)$.

Proof. In practical bibliographic databases, a subject classification scheme C consists of finite elements c_i that are mapped to finite subsets O_i using a map f. Let $\mathfrak{O}$ be a subset of $\mathfrak{P}(S)$ comprising $\{O_i|i\in I\}$, where I is a finite index set. For reference, let $\mathfrak{O}_0$ be the subset of $\mathfrak{P}(S)$ that consists of $\{\cap_{i\in I}A_i|A_i\in\mathfrak{O}\}$, where the element is S if $I=\varnothing$. Let $\widetilde{\mathfrak{O}}$ be a subset of $\mathfrak{P}(S)$ comprising $\{\cup_{\lambda\in\Lambda}B_\lambda|B_\lambda\in\mathfrak{O}_0\}$, where the element is $\varnothing$ if $\Lambda=\varnothing$. Here Λ is a finite or infinite index set. Thus, $\widetilde{\mathfrak{O}}\supset\mathfrak{O}$, $S\in\widetilde{\mathfrak{O}}$, and $\varnothing\in\widetilde{\mathfrak{O}}$, where $\widetilde{\mathfrak{O}}$ is satisfied as a topology using the necessary and sufficient conditions. Theorem 1 also indicates a compact topological space $\left(S,\widetilde{\mathfrak{O}}\right)$. When a finite cover exists in a topological space, we refer to it as a compact topological space.

Compact Topological Space Formation for a New Subject Classification Scheme

According to the subject classification model of the bibliographic database, we propose an approach of applying a new subject classification scheme for the database.

Here, we assume the following condition. A subject classification scheme $C^{(1)}$ containing subject categories $c_i^{(1)}$ is mapped to a finite cover $\mathfrak{O}^{(1)}=\left\{O_i^{(1)}\middle|i\in I^{(1)}\right\}$ using a map f_1, indicating a compact topological space $\left(S,\widetilde{\mathfrak{O}}^{(1)}\right)$.

Conventionally, we can use an approach to directly assign subject categories for the database records. We assign subject categories $c_i^{(2)}$ of a new classification scheme $C^{(2)}$ to each article of S. Thus, a map f_2 from $C^{(2)}$ to a finite cover $\mathfrak{O}^{(2)}=\left\{O_i^{(2)}\middle|i\in I^{(2)}\right\}$ is constructed, implying a compact topological space $\left(S,\widetilde{\mathfrak{O}}^{(2)}\right)$.

In our approach, we develop a correspondence $\Gamma:C^{(2)}\to C^{(1)}(\Gamma=\left(C^{(2)},C^{(1)};G\right),G\subset C^{(2)}\times C^{(1)})$, where $c_i^{(2)}\in C^{(2)},c_j^{(1)}\in C^{(1)}$, $c_i^{(2)}\times c_j^{(1)}\in G,C^{(2)}=\bigcup_i\left\{c_i^{(2)}\right\}$, and $C^{(1)}=\bigcup_j\left\{c_j^{(1)}\right\}$, to guarantee the existence of a finite cover.

Then, we construct a map

$$g_1:C^{(2)}\to\overline{\mathfrak{C}}^{(1)}=\left\{\overline{C}_i^{(1)}\middle|\begin{array}{c}c_i^{(2)}\in C^{(2)},\ c_j^{(1)}\in C^{(1)},\ c_i^{(2)}\times c_j^{(1)}\in G,\\ i\in I^{(2)},\ \overline{C}_i^{(1)}=\bigcup_{j\in I_i^{(1)}}\left\{c_j^{(1)}\right\}\end{array}\right\},$$

where $S = \bigcup_{i \in I^{(2)}} \overline{C}_i^{(1)}$, to be a finite cover. Finally, we establish a map

$$g_2 : \overline{\mathfrak{C}}^{(1)} \to \overline{\mathfrak{O}}^{(1)} = \left\{ \overline{O}_i^{(1)} \middle| \begin{array}{c} \overline{C}_i^{(1)} \in \overline{\mathfrak{C}}^{(1)}, c_j^{(1)} \in \overline{C}_i^{(1)}, O_j^{(1)} = f_1\left(c_j^{(1)}\right), \\ \overline{O}_i^{(1)} = \bigcup_{j \in I_i^{(1)}} O_j^{(1)} \end{array} \right\},$$

where $S = \bigcup_{i \in I^{(2)}} \overline{O}_i^{(1)}$, to be a finite cover. Hence, we obtain a composite map $g_2 \circ g_1$ from $C^{(2)}$ to a finite cover $\overline{\mathfrak{O}}^{(1)}$, indicating a compact topological space $\left(S, \widetilde{\overline{\mathfrak{O}}}^{(1)}\right)$. Evidently, $\widetilde{\overline{\mathfrak{O}}}^{(1)} \subset \widetilde{\mathfrak{O}}^{(1)}$.

Inducing a Correspondence Between Two Subject Classification Schemes Using a Research Project Database

To determine the correspondence between two subject classification schemes, experts of the subject classification schemes normally discuss the relationship structure of these schemes based on their knowledge and practical experiences.

In our approach, the actors are data scientists who analyze a database wherein an entity is categorized into two subject classification schemes and then induce the correspondence between them through an analysis.

As evidence data, anything that includes information of the relationship between the two subject classification schemes is useful. One of the available resources is a research project database that is rather popular among academic databases. In our case, it is the research project database KAKEN that includes the structural relationship between the WoS and KAKENHI subject categories. Thus, we ensure that our approach can adopt the research project database.

Using a Research Project Database

We define a research project database such as the KAKEN database as follows. A research project database T describes research projects b_n, one of whose outputs is a list of research articles a_n on a bibliographic database S.

The research articles a_n of S are categorized with a subject classification scheme $C^{(1)}$. We define the map f_1 by which $C^{(1)}$ is mapped to a finite cover $\mathfrak{O}_S^{(1)} = \left\{ O_i^{(1)} \middle| i \in I^{(1)} \right\}$ of S, implying a compact topological space $\left(S, \widetilde{\mathfrak{O}_S}^{(1)}\right)$.

The research projects b_n of T are categorized with a subject classification scheme $C^{(2)}$. We define a map h_1 by which $C^{(2)}$ is mapped to a finite cover $\mathfrak{O}_T^{(2)} = \left\{ O_i^{(2)} \middle| i \in I^{(2)} \right\}$ of T, implying a compact topological space $\left(T, \widetilde{\mathfrak{O}_T}^{(2)}\right)$.

We define a map $h_2 : T \to \mathfrak{P}(S)$ to ensure that a research project produces a set of research articles. Here, let the image of the map be reduced to $\mathfrak{S}$ $(\subset \mathfrak{P}(S))$ to become a surjection. Then, we also define a map $h_2' : T \to \mathfrak{P}(S')$, where $S' = \bigcup_{i \in I_{\mathfrak{S}}} O_i (O_i \in \mathfrak{S})$ and $S' \subset S$. For the image S', we define a map f_1' by which $C^{(1)}$ is mapped to a finite cover $\mathfrak{O}_{S'}^{(1)} = \left\{ O_i'^{(1)} \middle| i \in I^{(1)} \right\}$ of S', implying a compact topological space $\left(S', \widetilde{\mathfrak{O}_{S'}}^{(1)}\right)$.

Next, we develop a map

$$h_3 : \mathfrak{O}_T^{(2)} \to \mathfrak{O}_{S'}^{(2)} = \left\{ \overline{O}_{S'i}^{(2)} \middle| \begin{array}{c} O_{Ti}^{(2)} \in \mathfrak{O}_T^{(2)}, b_j^{(2)} \in O_{Ti}^{(2)}, O_{S'j}^{(2)} = h_2'\left(b_j^{(2)}\right), \\ \overline{O}_{S'i}^{(2)} = \bigcup_j O_{S'j}^{(2)} \end{array} \right\},$$

which is a subset of $\mathfrak{P}(S')$, where $\mathfrak{O}_{S'}^{(2)}$ is a finite cover. Subsequently, we obtain a composite map $h_3 \circ h_1 : C^{(2)} \to \mathfrak{O}_{S'}^{(2)}$. Considering that $\mathfrak{O}_{S'}^{(2)}$ is a finite cover, it induces a compact topological space.

In this case, we validated the following robust suppositions. The composite map $h_3 \circ h_1 : C^{(2)} \to \mathfrak{O}_{S'}^{(2)}$ represents the classification of articles using the subject classification scheme. Moreover, if two images on S' of maps f_1' and $h_3 \circ h_1$ are equivalent, then their inverse images also have the same relation.

Natural Overlapping Between Two Subject Classification Schemes

Here, we obtained actual data on the relationship between two subject classification schemes on a database. We have a database S' and two sets of finite covers $\mathfrak{O}_{S'}^{(1)}$ and $\mathfrak{O}_{S'}^{(2)}$ that are images from $C^{(1)}$ and $C^{(2)}$.

In natural phenomena, we often observe statistical laws of nature. A popular law in the linguistic field, that is, Zipf's law, states that the frequency of words follows a distribution where the word rank n has a frequency proportional to $1/n$. Generally, the same distribution is observed in natural phenomena, referred to as a power law, which is denoted by $\ln p(x) = -\alpha \ln x + c$, where α and c are constants (Newman, 2005). For example, all the following obey power law distributions: the sizes of city populations, earthquakes, moon craters, solar flares, computer files, and wars; the occurrence frequency of personal names in most cultures; the number of papers written by scientists; the number of citations received by papers; the number of hits on web pages; and the sales of books, music recordings, and almost every other branded commodity.

When actual data are analyzed, the power law trend in most cases holds only for an intermediate range of values; a power law breakdown exists in the distribution tails (Martínez-Mekler et al., 2009). The reason for this is finite size effects (e.g., insufficient data for good statistics), network dilution, network growth constraints, and different underlying dynamical regimes. Thus, power law corrections (sometimes referred to as scaling corrections) occur in the form of exponential, Gaussian, stretched exponential, gamma, and various types of extreme value distributions. This phenomenon obeys a discrete version of a generalized beta distribution, which is given by $f(r) = \left(A(N+1-r)^b\right)/r^a$. Here, r is the rank, N is its maximum value, A denotes the normalization constant, and (a, b) are two fitting exponents.

In our case, the elements of finite covers $\mathfrak{O}_{S'}^{(1)}$ and $\mathfrak{O}_{S'}^{(2)}$ represent natural overlapping sets. For $O^{(2)} (\in \mathfrak{O}_{S'}^{(2)})$, its

intersections $O^{(2)} \cap O^{(1)}$ to all $O^{(1)}(\in \mathfrak{O}^{(1)}_{S'})$ are present. Its cardinalities greater than zero, if sorted in rank order, obey the discrete version of the generalized beta distribution given that the subject categories are finite.

Metrics for Inducing a Correspondence

To identify a correspondence between $C^{(1)}$ and $C^{(2)}$, we attempt to find a subset $\left\{O^{(1)}_i \middle| i \in I^{(1)}_j\right\}$ of $\mathfrak{O}^{(1)}_{S'}$ for $O^{(2)}_{j\in I^{(2)}}$ to be ideally satisfied that $O^{(2)}_j = \bigcup_{i\in I^{(1)}_j} O^{(1)}_i$. However, in most cases, $O^{(2)}_j \not\supset O^{(1)}_i$ and $O^{(2)}_j \neq \bigcup O^{(1)}_i$. Hence, we first define the following metrics: (precision)

$$d_{pj} = \frac{\left|\bigcup_{i\in I^{(1)}_j}\left(O^{(2)}_j \cap O^{(1)}_i\right)\right|}{\left|\bigcup_{i\in I^{(1)}_j} O^{(1)}_i\right|}$$

and (recall)

$$d_{rj} = \frac{\left|\bigcup_{i\in I^{(1)}_j}\left(O^{(2)}_j \cap O^{(1)}_i\right)\right|}{\left|O^{(2)}_j\right|}.$$

Then, we define the generalized harmonic mean of precision and recall: (F_β-measure)

$$d_{fj} = \frac{\left(1+\beta^2\right) d_{pj} d_{rj}}{\beta^2 d_{pj} + d_{rj}},\ \beta > 0.$$

Finally, we use the F_β-measure to determine which element has a correspondence relation. The basic strategy is to choose the subset $\left\{O^{(1)}_i \middle| i \in I^{(1)}_j\right\}$ which maximize the F_β-measure. β affects the weight balance between d_{pj} and d_{rj} for d_{fj}. $\beta = 1$ indicates the equivalent balance between them. We can use the β to control the balance in relation to the existence of a finite cover.

In practical cases, we might project the cardinal number of the subsets onto the contingency table between two subject classification schemes. A contingency table, or a two-way frequency table, is a tabular mechanism with rows and columns used in statistics to present categorical data in terms of frequency counts.

By using the contingency table that represents the overall counting of elements, we calculate the following pseudo precision, recall, and F_β-measure based on the original definitions:(pseudo precision)

$$d'_{pj} = \frac{\sum_i \left|O^{(2)}_j \cap O^{(1)}_i\right|}{\sum_i \left|O^{(1)}_i\right|}$$

and (pseudo recall)

$$d'_{rj} = \frac{\sum_i \left|O^{(2)}_j \cap O^{(1)}_i\right|}{\left|O^{(2)}_j\right|}.$$

Then, the generalized harmonic mean of precision and recall is also calculated: (pseudo F_β-measure)

$$d'_{fj} = \frac{\left(1+\beta^2\right) d'_{pj} d'_{rj}}{\beta^2 d'_{pj} + d'_{rj}},\ \beta > 0.$$

The values of precision and pseudo precision, the values of recall and pseudo recall, and the values of F_β-measure and pseudo F_β-measure can be different because of subadditivity.

The Main Steps to Work on Our Approach

To follow the methodology above, the main steps of the work can be illustrated in **Figure 1**. The first step is to induce a correspondence between two subject classification schemes by using F_β-measure (step 1 in the figure). In practical cases, alternatively, the first step is to construct a contingency table between two subject classification schemes (step 1'-1) and then induce a correspondence between them by using pseudo F_β-measure (step 1'-2). The second step is to revise the correspondence to guarantee the existence of a finite cover of the novel subject classification scheme (step 2).

CASE STUDY

To verify our approach described previously, we adapt it for a practical case. A world-leading research output evaluation tool, that is, InCites™, which is produced by Clarivate Analytics, Co., Ltd., provides bibliometric analysis functions, wherein bibliometrics can be analyzed using domestic, WoS, and ESI subject classification schemes. Japanese users are eager to utilize the subject classification scheme of Japan's largest national research grants KAKENHI to analyze their IR outputs on the system. The WoS citation database comprises bibliographic records originally classified using the WoS subject classification scheme. The KAKENHI subject classification scheme is a novel subject classification scheme to be applied in the WoS citation database. We were occasionally given an opportunity to deal with this challenging task.

Inducing a Correspondence Between the WoS and KAKENHI Subject Categories

We use the following steps to induce a correspondence between the WoS and KAKENHI subject categories.

Developing a Contingency Table as Evidence Data

We construct a contingency table between the WoS and KAKENHI subject categories to induce a correspondence.

The research project database KAKEN represents the archival records of research projects and the outputs of KAKENHI grants in Japan. The KAKEN database contains the descriptions of projects started after 1964 and the lists of their outputs, including journal articles, conference proceedings, reports, and books. The research projects are classified using the KAKENHI subject classification scheme that has been defined for the corresponding year.

In this study, we select the research projects in 2009 whose KAKENHI subject classification scheme consists of a hierarchical

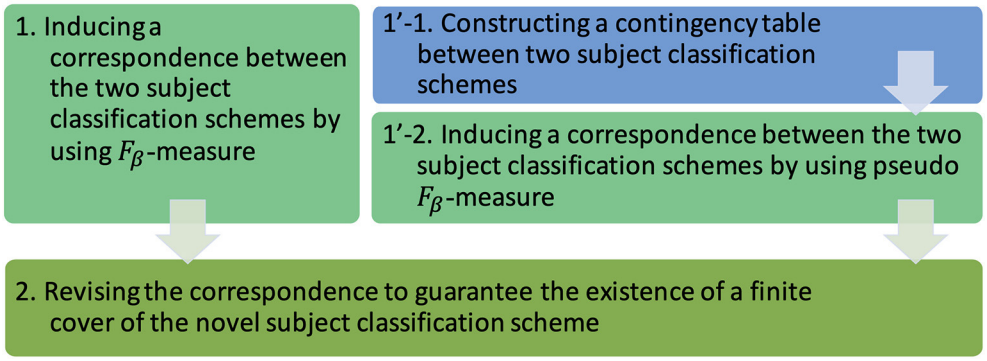

FIGURE 1 | The main steps to work on our methodology: constructing a contingency table, inducing a correspondence, and revising the correspondence.

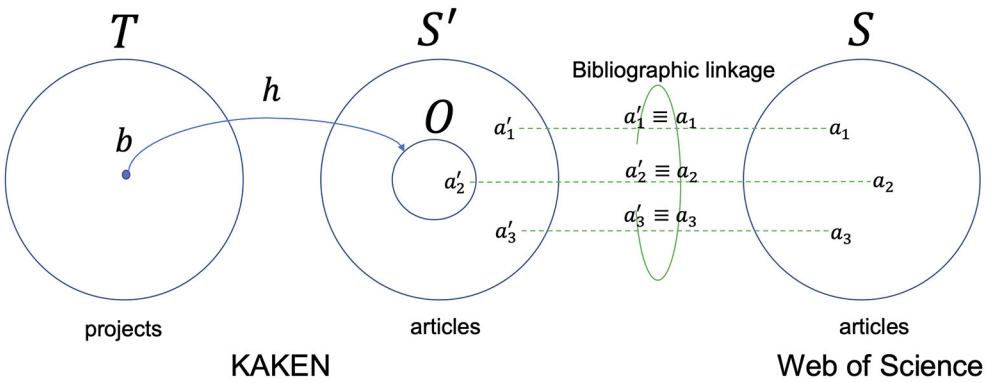

FIGURE 2 | Bibliographic linkage between the KAKEN and WoS citation databases in a Venn diagram.

structure: four categories, 10 areas, 67 disciplines, and 284 research fields. The total number of projects is 58,952, and that of output publications that might be written in English is 173,940.

The English articles in the KAKEN database are listed in a citation format, and it is not yet clear to which WoS categories they are assigned. Hence, we identified the same bibliographic records in the WoS citation database as of 2009 and 2010 (99.8% of output publications are published in the years) through a set of record linkage techniques to obtain a set of articles S' that are classified using both the KAKENHI and WoS classification schemes, as depicted in **Figure 2** (Kurakawa et al., 2014). The size of the adopted WoS citation database was 3,925,776, which is classified with 251 subject categories of the WoS classification scheme and 22 subject categories of the ESI classification scheme.

Consequently, we obtained a total of 75,042 pairs of citations, which is 43.1% of the 173,940 articles listed in the KAKEN database. The record linkage technique uses i-Linkage (Aizawa and Oyama, 2005) as a ranking function and SVM as a classification function to identify the same bibliographic records in the KAKEN database and the WoS citation database. In a 10-fold cross validation of 800 samples, the accuracy of the linkage was 0.9501. The precision, recall, and f-measure were 0.9492, 0.9510, and 0.9498, respectively.

We next constructed a contingency table for the two subject classification schemes based on this linkage result, as illustrated in **Figure 3**. An example in **Figure 4** shows part of the contingency table between the third-level 67 KAKENHI and 251 WoS subject categories.

Among the 75,042 pairs of citations, those categorized with both the subject classification schemes were reduced to 59,595 pairs because the 52,956 out of the total 58,952 research projects are assigned with the KAKENHI subject classification scheme and the others are not.

When the overall counting of the citations to each subject category was applied, we obtained the sum of 97,175 frequency counts in the contingency table. In the WoS citation database, each article is assigned one or more subject categories of the WoS classification scheme, and is simultaneously assigned one subject category of the ESI classification scheme. When we count a citation assigned to multiple subject categories, the frequency count is increased by one for each corresponding subject category. In the KAKEN database, each article is assigned one subject category of the KAKENHI classification scheme, the frequency count is increased by one for the corresponding subject category. Thus, for a citation, the frequency counts in the contingency table are increased by the number of WoS categories or ESI categories, and it looks

like many articles were published under the corresponding KAKENHI category.

Analysis of the Contingency Table

To clearly show what happens in the contingency table, we compared the distribution among the WoS subject categories with that of the KAKENHI subject category. We observed a good fit of the discrete generalized beta distribution to the rank-ordering distribution in the contingency table.

Figures 5, **6** show the rank-ordering distributions for the first and second levels of the subject categories of the KAKENHI subject classification scheme. The first-level subject categories include "Integrated Science and Innovative Science" (l1-01), "Humanities and Social Sciences" (l1-02), "Science and Engineering" (l1-03), and "Biological Sciences" (l1-04). The second-level subject categories include "Comprehensive Fields" (l2-01), "New Multidisciplinary Fields" (l2-02), "Humanities" (l2-03), "Social Sciences" (l2-04), "Mathematical and Physical Sciences" (l2-05), "Chemistry" (l2-06), "Engineering" (l2-07), "Biology" (l2-08), "Agricultural Sciences" (l2-09), and "Medicine, Dentistry, and Pharmacy" (l2-10). For each KAKENHI subject category at any level, the frequencies corresponding to the 251 WoS subject categories are sorted in rank order. If the frequency is zero, then the WoS subject category is omitted in the distribution. The x axis of the graph represents the rank, and the y axis of the graph denotes the log scale of the frequency count. With these scales, the discrete generalized beta distribution is fitted to the data to ensure that R-squared as a goodness-of-fit statistical score ranges from 0.986 to 0.994 for the first level and from 0.970 to 0.994 for the second level. In this case, the sets of parameters a and b that affect the figures of the distribution vary.

The distributions in the graph can be categorized into two types: concentration and dispersal. In the first level of the KAKENHI subject categories, the concentration type refers to the graph of "Science and Engineering" (l1-03) and "Biological Sciences" (l1-04). The dispersal type refers to the graph of "Integrated Science and Innovative Science" (l1-01). In the second level, the concentration type refers to

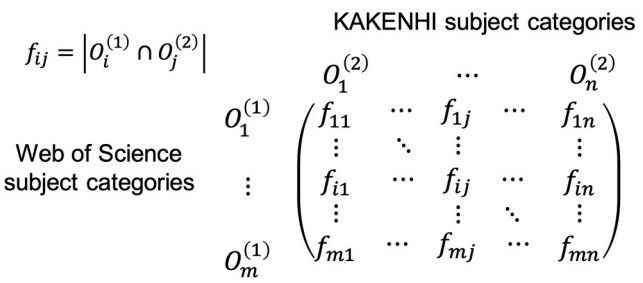

FIGURE 3 | Contingency table for the KAKENHI and WoS subject categories.

	l3-01	l3-02	l3-03	l3-04	l3-05	l3-06	l3-07	l3-08	l3-09	l3-10	l3-11	l3-12	l3-13	l3-14	l3-15	l3-16	l3-17	l3-18	l3-19	l3-20	l3-21	l3-22	l3-23
	情報学	神経科学	実験動物学	人間医工学	健康・スポーツ科学	生活科学	科学教育・教育工学	科学社会学・科学技術史	文化財科学	地理学	環境学	ナノ・マイクロ科学	社会・安全システム科学	ゲノム科学	生物分子科学	資源保全学	地域研究	ジェンダー	哲学	芸術学	文学	言語学	史学
Acoustics	59	0	0	10	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
Agricultural Economics & Policy	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Agricultural Engineering	0	0	0	0	0	0	2	0	0	0	4	0	0	0	0	1	2	0	0	0	0	0	0
Agriculture, Dairy & Animal Science	0	1	1	0	0	1	0	0	0	0	1	0	0	1	0	8	0	0	0	0	0	0	0
Agriculture, Multidisciplinary	0	0	0	0	0	4	0	0	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0
Agronomy	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	1	0	0	0	0	0	0
Allergy	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Anatomy & Morphology	0	11	0	1	4	1	0	0	0	0	0	0	0	3	2	0	0	0	0	0	0	0	0
Andrology	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Anesthesiology	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Anthropology	0	0	0	0	4	0	0	0	1	0	1	0	0	0	0	0	5	0	0	0	0	0	10
Archaeology	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Architecture	0	0	0	1	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Area Studies	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	4	2	0	0	1	1	4
Art	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	2	1	1	1	1
Asian Studies	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	1	0	1	0	10
Astronomy & Astrophysics	2	0	0	0	1	0	4	0	1	0	9	1	4	0	0	1	0	0	0	0	0	0	0
Audiology & Speech-Language Pathology	7	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	4	0
Automation & Control Systems	124	0	0	22	0	0	2	0	0	0	0	0	9	0	0	0	0	0	0	0	0	2	0
Behavioral Sciences	30	11	1	0	4	1	1	0	0	0	3	0	0	0	0	5	0	0	0	0	0	1	0
Biochemical Research Methods	31	15	1	17	0	2	0	0	2	0	10	20	0	15	10	0	0	0	0	0	0	0	0
Biochemistry & Molecular Biology	20	155	24	70	35	37	1	0	0	0	88	22	0	69	146	4	0	0	0	0	0	0	0
Biodiversity Conservation	0	0	0	0	0	0	2	0	0	0	24	0	0	0	0	22	0	0	0	0	0	0	0
Biology	8	6	3	6	9	1	0	0	0	0	35	2	1	6	0	2	0	0	1	0	0	0	0
Biophysics	12	27	3	31	14	5	0	3	0	0	22	14	0	5	16	0	0	0	2	0	0	0	0
Biotechnology & Applied Microbiology	31	8	4	39	2	15	3	1	2	0	67	10	1	21	37	2	3	0	0	0	0	0	0
Business	5	0	0	0	0	0	0	0	0	0	2	0	8	0	0	0	0	0	0	0	0	0	1
Business, Finance	4	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0

FIGURE 4 | Example screen of Excel showing part of the contingency table between the third-level subject categories of the KAKENHI and WoS subject classification schemes.

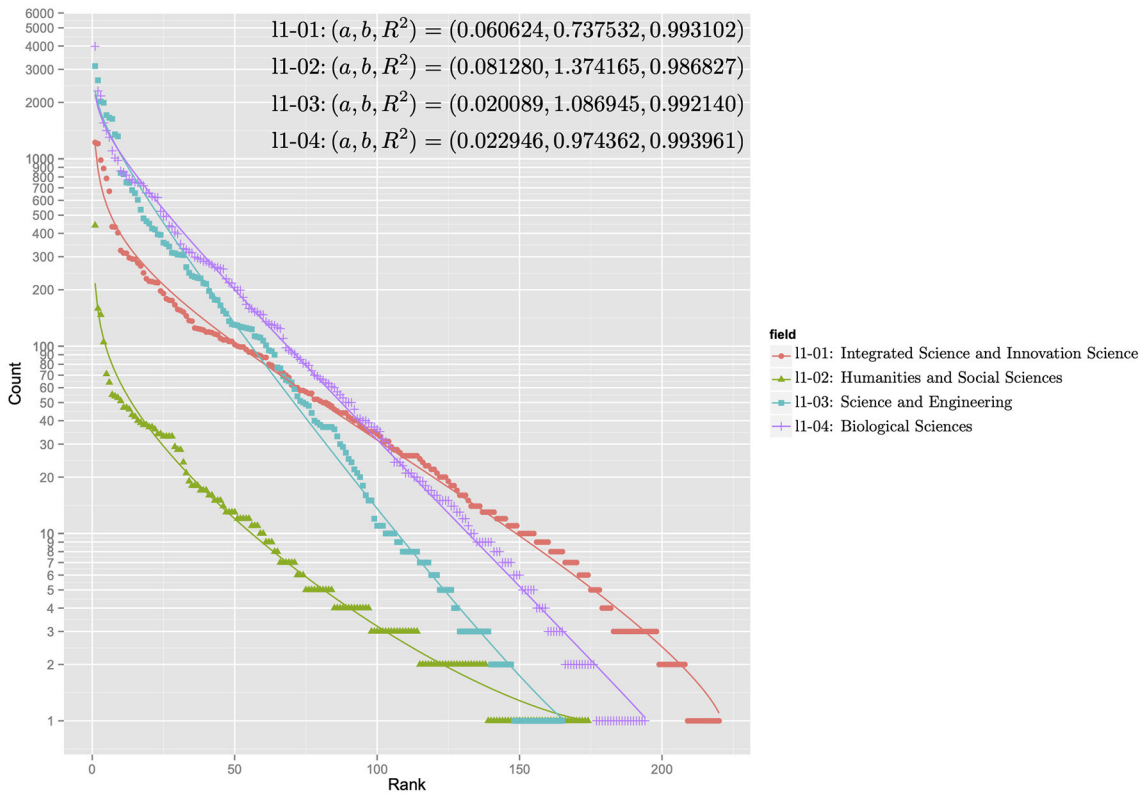

FIGURE 5 | Rank-ordering distributions for the first-level subject categories of the KAKENHI subject classification scheme.

the graph of "Humanities" (l2-03), "Chemistry" (l2-06), and "Mathematical and Physical Sciences" (l2-05). The dispersal type refers to the graph of "Comprehensive Fields" (l2-01) and "New Multidisciplinary Fields" (l2-02). The concentration type means the subject category is a specialized field. The dispersal type means the subject category is a multidisciplinary field.

For all the distributions, the goodness of fit to the discrete generalized beta distribution implies that a set of articles categorized to the KAKENHI subject category naturally overlaps that of articles categorized to the WoS subject categories at any level. However, the overlapping degree depends on the target subject categories.

Maximizing the F-Measure

Here, we analyzed the possibility that each KAKENHI subject category overlaps with the WoS subject categories. The aim of inducing a correspondence between the KAKENHI and WoS subject categories encouraged us to calculate the F_β-measures between these subject categories.

Table 1 lists the maximum pseudo F_1-measure, and the corresponding precision, and recall to produce the maximum pseudo F_1-measure, the number of WoS subject categories to get the maximum pseudo F_1-measure, and the cardinality for the third-level 67 disciplines of the KAKENHI subject classification scheme. The order of the disciplines in the list is the same as that of the KAKENHI subject classification scheme. The disciplines which are under the same area are listed together as a group. Each 67 discipline is included in either of 10 areas. For example, "Informatics" (l3-01), "Brain Sciences" (l3-02), "Laboratory Animal Science" (l3-03), "Human Informatics" (l3-04), "Health/Sports Science" (l3-05), "Human Life Science" (l3-06), "Science Education/Educational Technology" (l3-07), "Sociology/History of Science and Technology" (l3-08), "Cultural Assets Study" (l3-09), "Geography" (l3-10) are under the same area "Comprehensive Fields" (l2-01). In the same way, "Environmental Science" (l3-11) to "Gender" (l3-18) are under "New Multidisciplinary Fields" (l2-02), "Philosophy" (l3-19) to "Cultural Anthropology" (l3-25) are under "Humanities" (l2-03), "Law" (l3-26) to "Education" (l3-32) are under "Social Sciences" (l2-04), "Mathematics" (l3-33) to "Plasma Science" (l3-37) are under "Mathematical and Physical Sciences" (l2-05), "Basic Chemistry" (l3-38) to "Materials Chemistry" (l3-40) are under "Chemistry" (l2-06), "Applied Physics" (l3-41) to "Integrated Engineering" (l3-48) are under "Engineering" (l2-07), "Basic Biology" (l3-49) to "Anthropology" (l3-51) are under "Biology" (l2-08), "Plant Production and Environmental Agriculture" (l3-52) to "Boundary Agriculture" (l3-59) are under "Agricultural Sciences" (l2-09), "Pharmacy" (l3-60) to "Nursing" (l3-67) are under "Medicine, Dentistry, and Pharmacy" (l2-10).

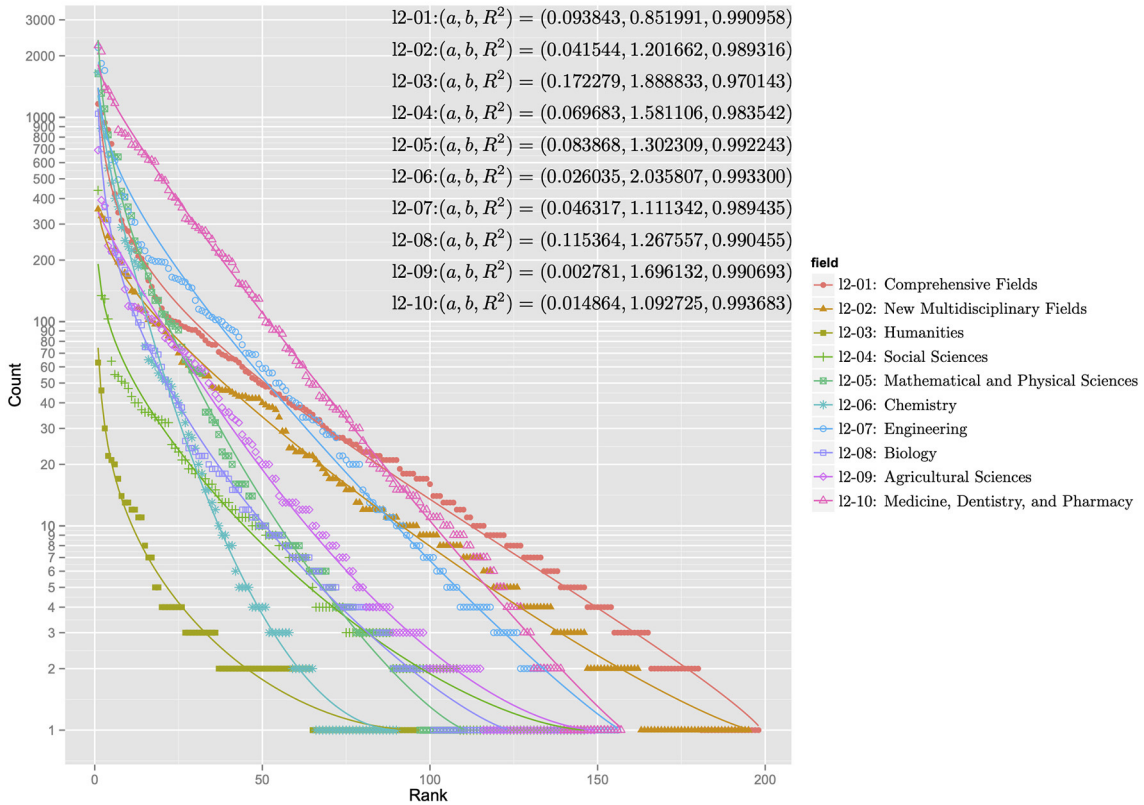

FIGURE 6 | Rank-ordering distributions for the second-level subject categories of the KAKENHI subject classification scheme.

Here, the top three subject categories which have greater maximum F_1-measure were "Mathematics" (l3-33), "Literature" (l3-21), and "Economics" (l3-28), whose pseudo average precision, recall, and maximum F_1-measure were (0.734, 0.792, 0.762), (0.700, 0.683, 0.691), and (0.692, 0.622, 0.655), respectively. The number of WoS subject categories to get the maximum pseudo F_1-measures were 4, 10, and 12, respectively. This means the corresponding WoS subject categories are much more relevant to the KAKENHI subject category. On the other hand, the bottom three subject categories which have less maximum F_1-measure were "Cultural Assets Study" (l3-09), "Laboratory Animal Science" (l3-03), and "Genome Science" (l3-14), whose pseudo average precision, recall, and maximum F_1-measure were (0.200, 0.036, 0.062), (0.059, 0.074, 0.066), and (0.040, 0.203, 0.067), respectively. The number of WoS subject categories to get the maximum pseudo F_1-measures were 1, 1, and 3, respectively. This means the corresponding WoS subject categories are less relevant to the KAKENHI subject category. The overall pseudo average precision, recall, and maximum F_1-measure were 0.315, 0.367, and 0.317, respectively. The number of WoS subject categories to get the maximum pseudo F_1-measures ranged from 1 to 24 whose average is 6.1, which is rather small when compared with the maximum number 251. The cardinality of the KAKENHI subject categories ranged from 9 to 10,215 whose average is 1,450.4. The larger the cardinality is, the more reliable the measure is, because of law of large numbers.

Miscellaneous Considerations

Apart from the quantitative analysis mentioned previously, we set the threshold in the contingency table to ignore the relations between the 251 WoS subject categories and the 67 disciplines of the KAKENHI subject categories. Here, for every WoS subject category $O_i^{(1)}$, the number of relations with the KAKENHI subject categories $O_j^{(2)}$ is only at a maximum of 1–4. Moreover, when the recall rate for $O_i^{(1)}$ exceeds 0.5, we discontinued adding any relation.

We next verified all the correspondence between $O_i^{(1)}$ and $O_j^{(2)}$ by means of subject category keywords, specifically for subject categories with few evidence data. The cases are "Arts and Humanities," "Music," and "Religion," among others of the $O_i^{(1)}$. This manual relation finding guarantees the existence of a finite cover by $O_j^{(2)}$.

Finally, we induced a correspondence between the 10 areas and 67 disciplines of the KAKENHI subject classification scheme and the 251 WoS subject categories, which are released in public[1]. A total of 324 relations are available in between the 10 areas of the KAKENHI subject classification scheme and the WoS subject categories, and 409 relations are present in between the 67 disciplines of

[1]Clarivate Analytics. *KAKEN Category Scheme - InCites Help*. Available online at: http://help.prod-incites.com/inCites2Live/filterValuesGroup/researchAreaSchema/kaken.html (accessed February 21, 2019).

TABLE 1 | Maximum pseudo F_1-measure for the third-level 67 disciplines of the KAKENHI subject categories against the 251 WoS subject categories.

The third-level 67 disciplines seq. no.	KAKENHI subject category	Translation	Cardinality	No. of WoS subject categories to get the max pseudo F_1-measure	Pseudo precision	Pseudo recall	Max. pseudo F_1-measure
(I3-01)	情報学	Informatics	6,637	17	0.576	0.626	0.600
(I3-02)	神経科学	Brain sciences	1,570	1	0.218	0.365	0.273
(I3-03)	実験動物学	Laboratory animal science	242	1	0.059	0.074	0.066
(I3-04)	人間医工学	Human informatics	1,995	8	0.222	0.213	0.217
(I3-05)	健康・スポーツ科学	Health/sports science	844	5	0.181	0.290	0.223
(I3-06)	生活科学	Human life science	467	4	0.239	0.281	0.258
(I3-07)	科学教育・教育工学	Science education/educational technology	388	2	0.377	0.103	0.162
(I3-08)	科学社会学・科学技術史	Sociology/history of science and technology	43	6	0.111	0.163	0.132
(I3-09)	文化財科学	Cultural assets study	55	1	0.200	0.036	0.062
(I3-10)	地理学	Geography	148	4	0.117	0.203	0.149
(I3-11)	環境学	Environmental science	2,136	14	0.262	0.385	0.312
(I3-12)	ナノ・マイクロ科学	Nano/micro science	1,852	4	0.103	0.313	0.155
(I3-13)	社会・安全システム科学	Social/safety system science	868	14	0.187	0.214	0.199
(I3-14)	ゲノム科学	Genome science	394	3	0.040	0.203	0.067
(I3-15)	生物分子科学	Biomedical engineering	875	2	0.119	0.325	0.174
(I3-16)	資源保全学	Culture assets and museology	172	3	0.181	0.145	0.161
(I3-17)	地域研究	Area studies	85	7	0.164	0.271	0.204
(I3-18)	ジェンダー	Gender	27	3	0.231	0.111	0.150
(I3-19)	哲学	Philosophy	60	4	0.436	0.283	0.343
(I3-20)	芸術学	Art Studies	9	1	0.091	0.111	0.100
(I3-21)	文学	Literature	41	10	0.700	0.683	0.691
(I3-22)	言語学	Linguistics	239	3	0.705	0.410	0.519
(I3-23)	史学	History	82	6	0.412	0.341	0.373
(I3-24)	人文地理学	Human geography	14	3	0.175	0.500	0.259
(I3-25)	文化人類学	Cultural anthropology	38	3	0.056	0.105	0.073
(I3-26)	法学	Law	41	3	0.385	0.122	0.185
(I3-27)	政治学	Politics	59	2	0.409	0.458	0.432
(I3-28)	経済学	Economics	992	12	0.692	0.622	0.655
(I3-29)	経営学	Management	130	5	0.294	0.385	0.333
(I3-30)	社会学	Sociology	90	8	0.176	0.278	0.216
(I3-31)	心理学	Psychology	794	14	0.488	0.479	0.483
(I3-32)	教育学	Education	151	9	0.244	0.258	0.251
(I3-33)	数学	Mathematics	2,589	4	0.734	0.792	0.762
(I3-34)	天文学	Astronomy	1,005	1	0.505	0.870	0.639
(I3-35)	物理学	Physics	5,199	6	0.498	0.651	0.565
(I3-36)	地球惑星科学	Earth and planetary science	2,099	7	0.619	0.662	0.640
(I3-37)	プラズマ科学	Plasma science	508	1	0.233	0.191	0.210
(I3-38)	基礎化学	Basic chemistry	2,448	7	0.229	0.801	0.356
(I3-39)	複合化学	Applied chemistry	3,573	6	0.283	0.526	0.368
(I3-40)	材料化学	Materials chemistry	1,635	7	0.157	0.348	0.216
(I3-41)	応用物理学・工学基礎	Applied physics	2,235	5	0.170	0.394	0.238
(I3-42)	機械工学	Mechanical engineering	2,675	11	0.431	0.388	0.408

(Continued)

TABLE 1 | Continued

The third-level 67 disciplines seq. no.	KAKENHI subject category	Translation	Cardinality	No. of WoS subject categories to get the max pseudo F_1-measure	Pseudo precision	Pseudo recall	Max. pseudo F_1-measure
(I3-43)	電気電子工学	Electrical and electric engineering	4,875	10	0.338	0.669	0.449
(I3-44)	土木工学	Civil engineering	711	8	0.371	0.484	0.420
(I3-45)	建築学	Architecture and building engineering	170	3	0.286	0.506	0.365
(I3-46)	材料工学	Material engineering	2,931	6	0.348	0.523	0.418
(I3-47)	プロセス工学	Process/chemical engineering	1,283	4	0.145	0.306	0.197
(I3-48)	総合工学	Integrated engineering	1,465	8	0.256	0.309	0.280
(I3-49)	基礎生物学	Basic biology	2,423	7	0.375	0.400	0.387
(I3-50)	生物科学	Biological science	2,679	4	0.167	0.582	0.259
(I3-51)	人類学	Anthropology	300	3	0.315	0.440	0.367
(I3-52)	農学	Plant production and environmental agriculture	899	4	0.307	0.449	0.365
(I3-53)	農芸化学	Agricultural chemistry	1,755	6	0.220	0.386	0.281
(I3-54)	林学	Forest and forest products science	559	5	0.408	0.252	0.312
(I3-55)	水産学	Applied aquatic science	581	2	0.419	0.327	0.367
(I3-56)	農業経済学	Agricultural science in society and economy	31	2	0.333	0.097	0.150
(I3-57)	農業工学	Agro-engineering	216	4	0.157	0.259	0.195
(I3-58)	畜産学・獣医学	Animal life science	1,190	4	0.511	0.387	0.440
(I3-59)	境界農学	Boundary agriculture	541	4	0.235	0.148	0.181
(I3-60)	薬学	Pharmacy	3,457	4	0.294	0.369	0.328
(I3-61)	基礎医学	Basic medicine	5,232	16	0.213	0.551	0.307
(I3-62)	境界医学	Boundary medicine	850	12	0.162	0.112	0.132
(I3-63)	社会医学	Society medicine	1,065	8	0.282	0.262	0.271
(I3-64)	内科系臨床医学	Clinical internal medicine	10,215	24	0.441	0.617	0.514
(I3-65)	外科系臨床医学	Clinical surgery	5,562	20	0.418	0.468	0.442
(I3-66)	歯学	Dentistry	2,523	3	0.640	0.280	0.389
(I3-67)	看護学	Nursing	158	2	0.737	0.443	0.553
Average			1,450.4	6.1	0.315	0.367	0.317

the KAKENHI subject classification scheme and the WoS subject categories.

Classification Results on the WoS Citation Database

With the correspondence, InCites™ preprocesses its internal database and provides the analysis functionality by using the KAKENHI subject classification scheme. The techniques used by the tool in providing the analysis function, its quantitative statistics, and user feedbacks of the function are discussed in the following sections.

KAKENHI Subject Categories of InCites™

InCites™ provides an analytical workbench on the WoS citation database. It preprocesses the database to demonstrate users' target entities such as people, organizations, regions, research areas, journals, books, conference proceedings, and funding agencies. **Figure 7** shows an example screen presenting the article counts of Japanese authors based on the 67 disciplines of the KAKENHI subject classification scheme. The bubbles in the figure represent the top 25 proportional numbers of articles, each of which corresponds to the KAKENHI subject category. The total number of articles by the Japanese authors is 3,192,449 of the overall 58,395,008 articles published from 1980 to 2018. Among this Japanese authorship, the top or first KAKENHI subject category at the discipline level is "Clinical internal medicine," with a total number of 1,096,040. The second and third are "Basic medicine" and "Applied chemistry," with a total number of 617,970 and 526,139, respectively.

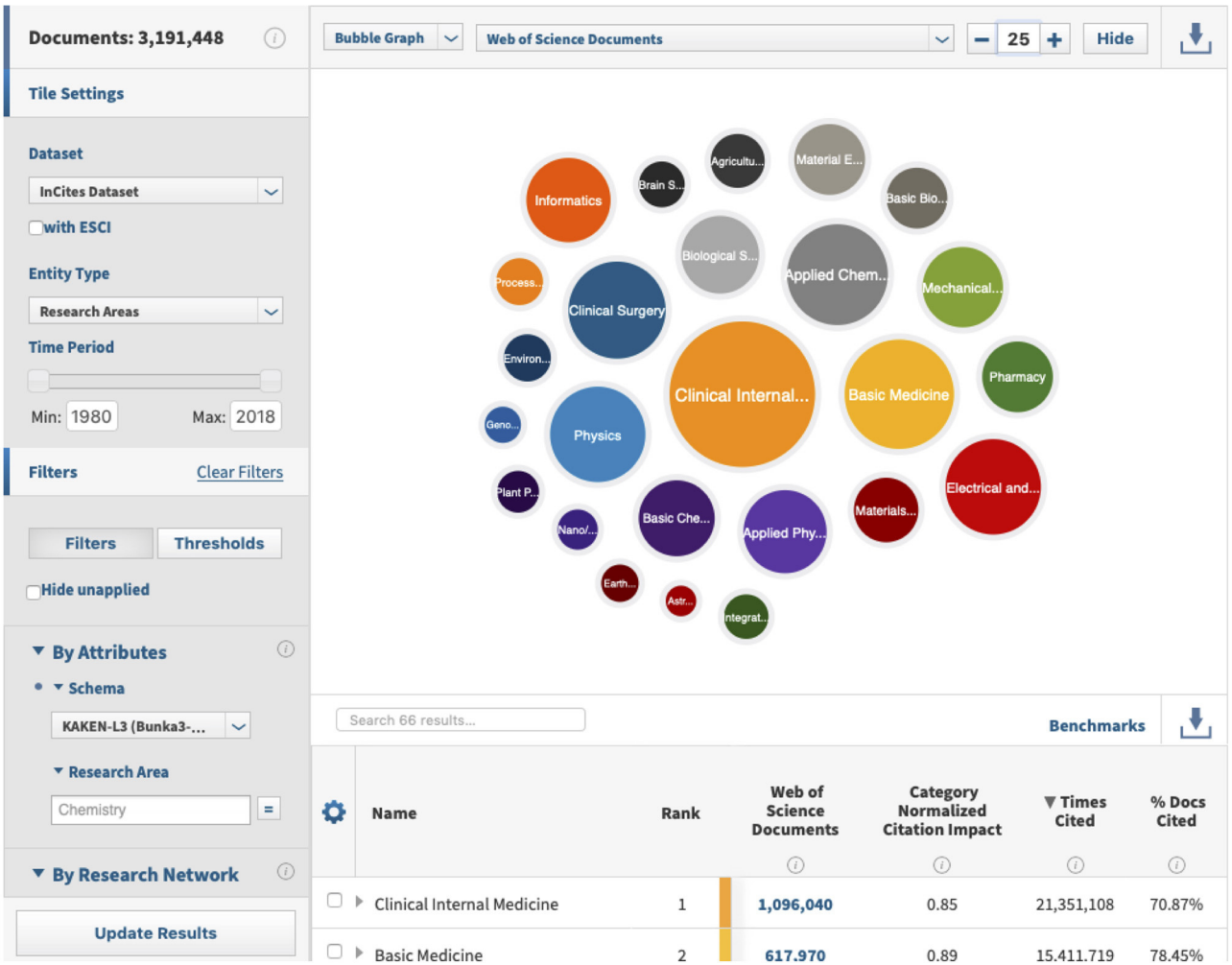

FIGURE 7 | Example screen of InCites™, with the bubbles representing proportional numbers of articles classified using the KAKENHI subject categories.

Article Counts Using the WoS and KAKENHI Subject Categories

For Japanese authors' articles, we compared the distributions using the subject classification schemes. We illustrated the proportions based on the statistics provided by the tool through the subject classification schemes shown in **Figures 8–10**.

Figure 8 shows the top 30 subject distribution of articles using the WoS subject classification scheme. At the top of the list are "Engineering, Electrical & Electronic," "Physics, Applied," "Biochemistry & Molecular Biology," "Materials Science, Multidisciplinary," and "Chemistry, Multidisciplinary," among others. The distribution of the graph gradually declines similar to an inverse proportional graph.

Figure 9 shows the subject distribution of the same set of articles with the 10 areas level of the KAKENHI subject classification scheme. At the top of the list are "Medical/Dental/Pharmaceutical," "Engineering," "Math/Physics," "Multidisciplinary," and "Chemistry," among others. The number of articles for the subject categories declines linearly rather than inversely. **Figure 10** shows the top 30 subject distribution of articles by the 67 disciplines level of the KAKENHI subject classification scheme. At the top of the list are "Clinical Internal Medicine," "Basic Medicine," "Applied Chemistry," "Clinical Surgery," and "Electrical and Electric Engineering," among others. The number of articles declines inversely. Unlike the original WoS subject categories, this statistical result provides a different impression that life sciences are the strongest among the others. However, the WoS subject classification scheme generates an impression that "Electrical/Electronic Engineering" and "Physics" are the strongest among the others.

User Feedback

In response to the KAKENHI subject classification scheme considering that a new function of InCites™ was released on April 2016, users in Japan were surveyed via an online questionnaire after a year: that is, in April 2017.

A total of 26 institutional users answered the questionnaire. They were mostly research administrators (RAs) and IR staff (**Table 2**).

The questionnaire comprises 18 questions related to the subject classification schemes implemented in InCites™ and the

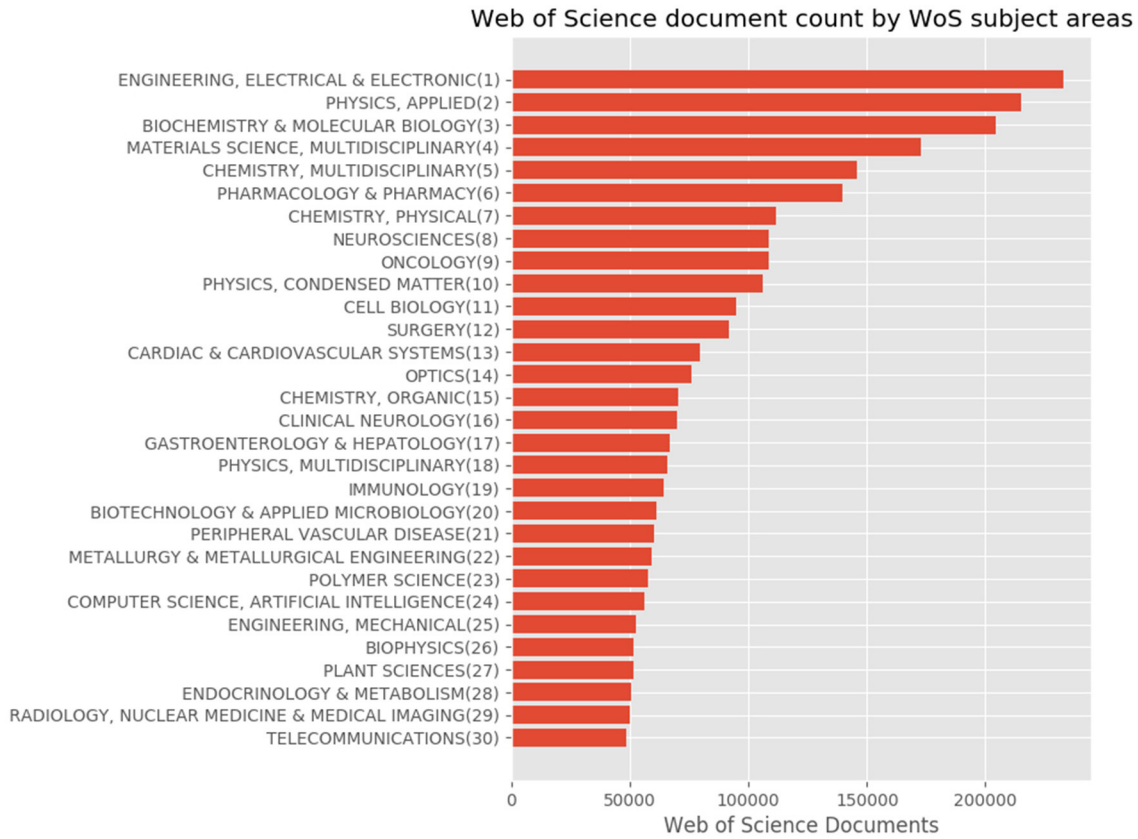

FIGURE 8 | Top 30 subject distribution of Japanese authors' articles by using the WoS subject classification scheme.

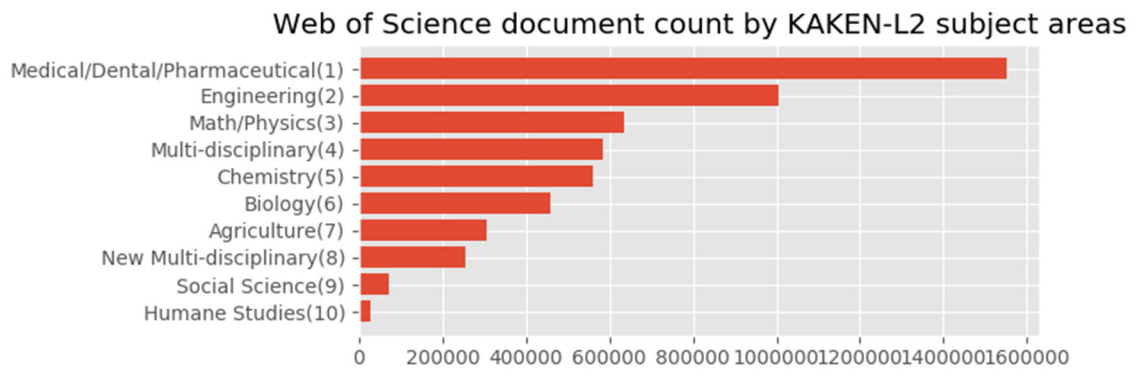

FIGURE 9 | Entire subject distribution of Japanese authors' articles with the 10 areas level of the KAKENHI subject classification scheme.

attributes of users. An open-ended question was provided in the final part of the questionnaire.

To determine the users' degree of expertise, Q13 and Q3 were prepared. Q13 asks how often users utilize InCites™, whereas Q3 asks how broad the users' knowledge is regarding the KAKENHI subject classification scheme. The results indicate that most of the users periodically utilize the tool in their work and have sufficient expertise on the KAKENHI subject classification scheme (**Figure 11**).

With regard to the validity of the KAKENHI subject classification scheme, Q7 and Q11 were asked. Q7 investigates the necessary hierarchy level of the KAKENHI subject classification scheme. Q11 asks whether the users are comfortable with the analysis results when they use the KAKENHI subject classification scheme. The results of these questions indicate that

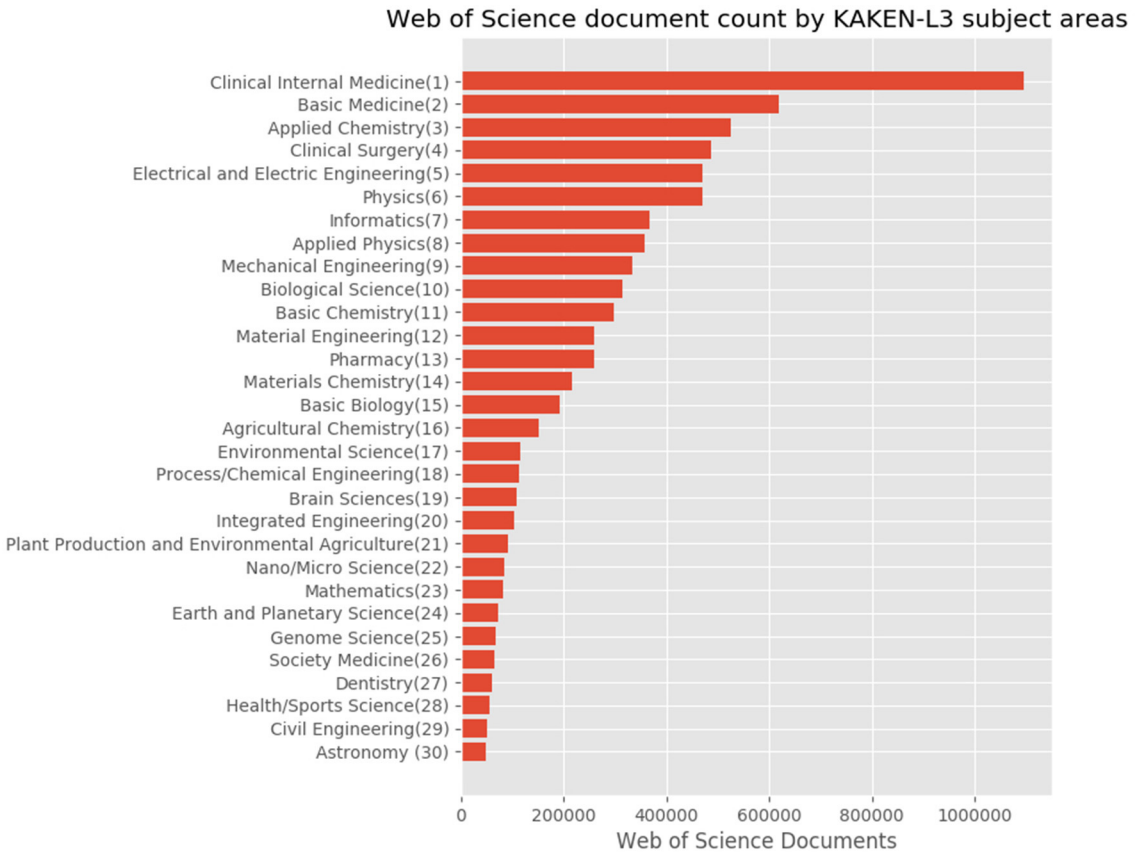

FIGURE 10 | Top 30 subject distribution of Japanese authors' articles with the 67 disciplines level of the KAKENHI subject classification scheme.

TABLE 2 | Users' role in their institutions.

Users' role in the institution	Yes (multiple answers possible)
RA (research administrator)	20
Administrator/officer	3
IR (institutional research) staff	5
Others	2

users think they require both levels of hierarchy and are almost satisfied with the analysis results by using the KAKENHI subject classification scheme when compared with their experience with KAKENHI funding-related jobs (**Figure 12**).

In the section of the questionnaire where further comments on the new feature of the KAKENHI subject classification scheme were encouraged, several users insisted on its usefulness. Moreover, users stated that they needed the same subject categories for other services and wanted them updated. Their exact comments were as follows:

- *"I need the KAKENHI subject classification scheme in the Web of Science search service as well."*
- *"I hope for updating the KAKENHI subject classification scheme to new* one *as possible. (It might be hard to catch up on updating it since it changes every year)."*
- *"Sixty-over categories of KAKENHI is not sufficient to relatively compare researches as much as ESI (22 only) and WoS (251,* four *times and more). And it may cause over-evaluation in comparison between research fields because the KAKENHI subject classification is made in a clock counter-like classification method. We need more accurate analysis of more concrete examples."*

Discussion

By analyzing the theory of our approach, that is, inducing a correspondence between the two subject classification schemes, we recognized its inherent limitation. The embedding subject classification scheme is unavoidably dependent on the original classification scheme. The topological space of the former is a subset of the topological space of the latter. However, we observed that each subject category of one scheme partially overlaps several subject categories of the other scheme based on the natural correlations between the subject categories of two subject classification schemes. No inclusion relationship exists between them. Therefore, the correspondence relations must be probabilistic.

In addition, we set strong assumptions on the relation among the research projects and journal articles in the research project database, in that they have similarities on the subject. However,

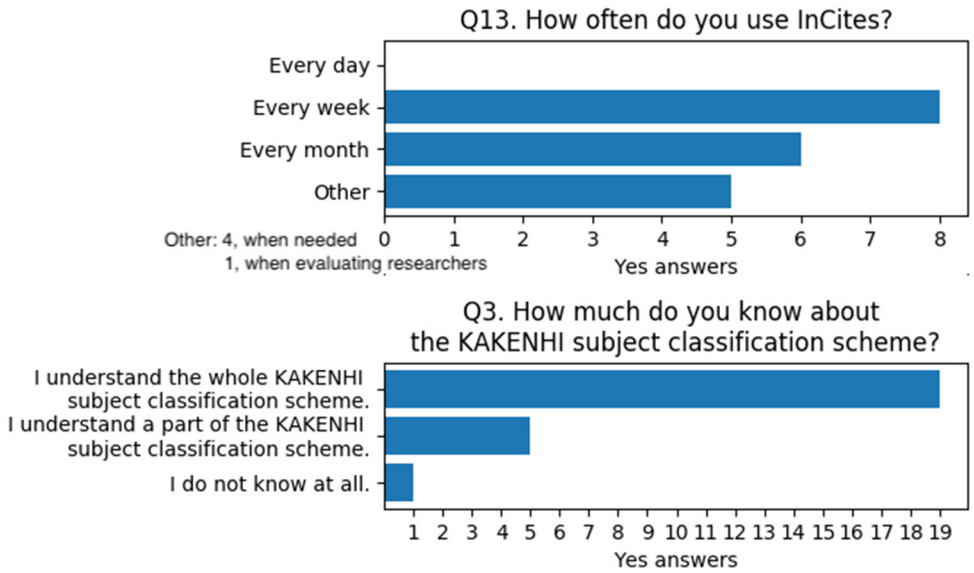

FIGURE 11 | Questions and answers on users' degree of expertise for InCites™ and the KAKENHI subject classification scheme.

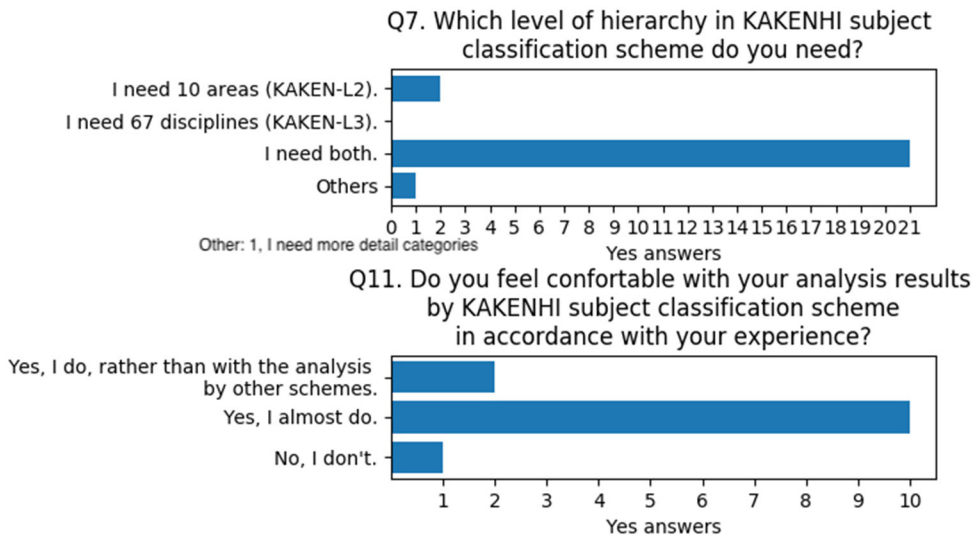

FIGURE 12 | Questions and answers on the validity of the KAKENHI subject classification scheme.

the research projects and article outputs do have both similarities and differences on the subject. On the similarity side, we used the following procedure. A grants database reports that research projects produce outputs, that is, research articles. We focused on the subject classification scheme for the research projects and its relationship to a set of research articles. These articles were classified by using another subject classification scheme. Then, we compared the relationship of these two subject classification schemes. On the difference side, we have another issue. For example, the projects precede the articles. The time lag between project initiation and article outputs is observed, rendering a subject divergence or drift between them. Moreover, the projects tend to indicate the central concept using important keywords, allowing a subject diversification of the articles.

Nevertheless, users of InCites™ accepted the subject classification results. We assumed the following reasons. First, users might focus on the comparative analysis of bibliometrics based on the subject categories and not care about the specific case of articles. Second, they might require a rough quality of metrics during the evaluation stage. Metrics are the central limits of the quantitative attributes of a set of entities, which is the main indicator to be verified during the research evaluation.

Another advantage is that our approach requires less workload. In 2019, the number of WoS documents stored in

InCites™ is 58,395,008, wherein the total number of journal titles amounts to 24,688. Thus, far, the possible targets for assigning subject categories are the WoS documents and journal titles. The journal titles include a set of documents. Furthermore, assigning subject categories to journal titles implies subsequently assigning them to the documents. In production, the WoS subject categories are primarily and exceptionally assigned to journal titles and documents in multidisciplinary journals. In our approach, we induced the correspondence between the WoS and KAKENHI subject classification schemes by using the KAKEN database. For the 251 WoS subject categories and 67 disciplines of the KAKENHI subject categories, the maximum relations in the correspondence are up to 16,817 (251×67). Regarding the 10 areas of KAKENHI subject categories, the maximum relations are up to 2,510 (251×10). The number for verifying the relations in our approach is overwhelmingly smaller than that of the original subject category manual assignment approach.

The evidence data are the contingency table whose sum of the frequency counts is 97,175. Specifically, this number is not sufficient for automatic decision making. When we assessed the correspondence between both subject classification schemes, the absence of relations is evident. The relations should be present in the literary meaning. Hence, manual handling was necessary for several subject categories. If the data size is sufficiently large, then we could predict the correspondence by using the data only.

CONCLUSIONS AND FUTURE WORK

In this study, we proposed an approach to apply a new subject classification scheme for a bibliographic database that is already classified by using a subject classification scheme. We also defined the subject classification model of the bibliographic database comprising a topological space. Then, we presented our approach based on the model, wherein forming a compact topological space is necessary for a novel subject classification scheme. To form the space, the correspondence between the two subject classification schemes by using the research project database was utilized as data.

We applied the approach to a practical example, that is, InCites™. This tool is used as a world proprietary benchmarking tool for research evaluation based on the WoS citation database to add the subject classification scheme of Japan's largest national grants KAKENHI. Finally, InCites™ provides a function of analysis by using the KAKENHI subject classification scheme. The survey indicates that users generally accept the new feature.

In future work, several aspects are necessary to improve the quality of the database and embed subject classification schemes by using effective and efficient automatic procedures. In real cases, there exists a more complex subject classification scheme. Our approach assumes that the subject classification schemes consist of a flat formation. For a complex classification scheme such as a hierarchical classification scheme, our approach should be extended to be applied to its character. Alternatively, multilabel learning is another possible method to aim at our goal. A comparative study is needed to qualify our method.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

This article is a result of a joint research between National Institute of Informatics and Clarivate Analytics, Co., Ltd. As for the databases we used in this article, the KAKEN database is provided by Scholarly and Academic Information Division, Cyber Science Infrastructure Development Department, National Institute of Informatics, and the Web of Science citation database is provided by Clarivate Analytics, Co., Ltd. We are thankful to the organizations who let us use the valuable assets.

REFERENCES

Aizawa, A., and Oyama, K. (2005). "A fast linkage detection scheme for multi-source information integration," in *International Workshop on Challenges in Web Information Retrieval and Integration* (Tokyo: IEEE), 30–39.

Bairi, R. B., Carman, M. J., and Ramakrishnan, G. (2016). "Beyond clustering," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16* (New York, NY: ACM Press), 801–810.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Büyükçakir, A., Bonab, H., and Can, F. (2018). "A novel online stacked ensemble for multi-label stream classification," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management - CIKM '18* (New York, NY: ACM Press), 1063–1072.

Chang, X., Yu, Y.-L., and Yang, Y. (2017). "Robust top- k multiclass SVM for visual category recognition," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17* (New York, NY: ACM Press), 75–83.

Gómez, I., Bordons, M., Fernández, M. T., and Méndez, A. (1996). Coping with the problem of subject classification diversity. *Scientometrics* 35, 223–235.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining : Concepts and Techniques*. Waltham, MA: Morgan Kaufmann, an imprint of Elsevier.

Hjørland, B. (2008). What is knowledge organization (KO)? *Knowl. Organ.* 35, 86–101. doi: 10.5771/0943-7444-2008-2-3-86

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Comput. Surv.* 31, 264–323. doi: 10.1145/331499.331504

Koch, T., Day, M., Brümmer, A., Hiom, D., Peereboom, M., Poulter, A., et al. (1997). *The Role of Classification Schemes in Internet Resource Description and Discovery*. Bath: UKOLN, University of Bath.

Kurakawa, K., Sun, Y., and Aizawa, A. (2014). Mapping Between Research Fields of Grants-in-Aid for Scientific Research and Web of Science Subject Areas. NII Technical Reports, National Institute of Informatics. Available online at: https://www.nii.ac.jp/TechReports/public_html/14-002J.html (accessed August 31, 2019).

Liu, J., Chang, W.-C., Wu, Y., and Yang, Y. (2017). "Deep learning for extreme multi-label text classification," in *Proceedings of the 40th International ACM*

SIGIR Conference on Research and Development in Information Retrieval - SIGIR '17 (New York, NY: ACM Press), 115–124.

Mai, F., Galke, L., and Scherp, A. (2018). "Using deep learning for title-based semantic subject indexing to reach competitive performance to full-text," in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries - JCDL '18* (New York, NY: ACM Press), 169–178.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Maron, M. E. (1961). Automatic indexing: an experimental inquiry. *J. ACM* 8, 404–417.

Martínez-Mekler, G., Alvarez Martínez, R., Beltrán del Río, M., Mansilla, R., Miramontes, P., and Cocho, G. (2009). Universality of rank-ordering distributions in the arts and sciences. *PLoS ONE* 4:e4791. doi: 10.1371/journal.pone.0004791

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* 46, 323–351. doi: 10.1080/00107510500052444

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.* 34, 1–47. doi: 10.1145/505282.505283

Slavic, A. (2011). "Classification revisited: a web of knowledge," in *Innovations in Information Retrieval: Perspectives for Theory and Practice*, eds A. Foster and P. Rafferty (London: Facet Publishing), 23–48.

Tagami, Y. (2017). "AnnexML," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17* (New York, NY: ACM Press), 55–464.

Wang, P., Yang, Z., Niu, S., Zhang, Y., Zhang, L., and Niu, S. (2018). "Modeling dynamic pairwise attention for crime classification over legal articles," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '18* (New York, NY: ACM Press), 485–494.

Xie, C., Chen, L., Liang, J., Zhang, K., Xiao, Y., Tong, H., et al. (2017). "Automatic navbox generation by interpretable clustering over linked entities," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17* (New York, NY: ACM Press), 1857–1865.

Conflict of Interest: The authors declare that this study received funding from Clarivate Analytics, Co., Ltd. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.