Vol.2019-IFAT-134 No.7
Vol.2019-DC-112 No.7
2019/3/22

Applying a New Subject Classification Scheme for a Database by a Data-driven Correspondence

KEI KURAKAWA †1 YUAN SUN †1
SATOKO ANDO †2

Abstract: In research evaluation, bibliographic database records are classified with a variety of subject classification schemes to be analyzed from various viewpoints. A new subject classification scheme often needs to be applied to a pre-classified bibliographic database for the evaluation task. Generally speaking, applying a new subject classification scheme is labor-intensive and time-consuming. It requires cost effective and efficient approach. So, we propose an approach to apply a new subject classification scheme for a subject classified database by a data-driven correspondence between the new and preset ones. In this paper, we define a subject classification model of database that consists of a topological space. Then, we show our approach based on the model, where the step is to form a compact topological space for a new subject classification scheme. To form the space, it utilizes a correspondence between two subject classification schemes by a research project database as data. For the case study of our approach, we applied it to a practical example, i.e. InCites™ - a world proprietary benchmarking tool for research evaluation based on the Web of Science citation database so as to add the subject classification scheme of Japan's national biggest grants KAKENHI. By means of the KAKEN database that keep records of research project descriptions and achievement lists of KAKENHI and record linkage techniques, i.e., i-Linkage and SVM, 59,595 pairs of articles classified with the both subject classification schemes are extracted. Then, we analyzed the pairs of articles so as to induce the correspondences between the 10 areas / 67 disciplines of KAKENHI subject categories and 251 Web of Science subject categories based on our approach. The InCites™ gives a function of analysis with the KAKENHI subject classification scheme. By a user survey, it is revealed that the users accepted the feature on the whole.

Keywords: Bibliographic database, Subject classification scheme, Topological space, Research project database, Data-driven correspondence

1. Introduction

Subject classification is a popular and useful aspect of academic database and data analysis. Academic resources such as research articles, journals, conference proceedings, books, samples in the field, software and a variety of electronic-materials are organized by subject classifications, which is in general or in domain-specific. University libraries, institutional resource centers as well as research labs organize their research resources in a good manner in order to get easily access to them on demand. Academic funding organizations manage their applicants, projects and reports by subject classifications of research, which is often diversified and transformed to reflect the current research landscape. Academic fields are fundamental concepts of academic classifications to organize academic materials. In analysis viewpoints, institutional research focus on research and educational activities, in which research and educational portfolios of researchers, professors, and staffs to be analyzed by subject classifications. National grants databases are often surveyed by the subject classifications.

Subject classification is a knowledge structure that is built in information science methodologies. To deal with information resources, we have two axes of objectives, i.e. library and knowledge. Library is a part of information science that aims at information management and knowledge organization, and for further information retrieval. It is supposed to deal with chunks of information and knowledge. Knowledge is another part of information science that is oriented to knowledge representation and extraction. Each of them gives two types of methods to accomplish its objectives, i.e. categorization method and terms-and-associations method. The categorization method represents classes of objects and defines inclusion relationships among them. The terms-and-associations method represents terms for classes of objects and defines a variety of associations between them. In library domain, the former is *classification*. The latter is *thesaurus*. In knowledge domain, the former is *taxonomy*. The latter is *ontology*.

In practice, classification has been utilized in library catalogues for a hundred years or so [1]. Dewey decimal classification (DDC) is the old library classification invented in 1876, which is popular to classify books in shelves of university libraries. The other popular library classifications such as universal decimal classification (UDC), library of congress classification (LCC), colon classification (CC) are also invented hundred years ago, and revised many times to fit a present book subject diversity until now. Japanese library classification examples are Nippon decimal code (NDC) and NDL classification (NDLC), which was released respectively in 1928 and 1963. For academic journals, Web of Science subject classification is well known for one of the subject classifications for the Web of Science citation database. For the research evaluation purpose, journals are accustomed to being classified from the specific viewpoints. Essential science indicator (ESI) subject classification is a representative that is essentially created for the research evaluation based on the Web of Science citation database.

In the domain of research evaluation, there exist strong needs to adopt special subject classifications that are created in research and educational works [2]. National research and educational

†1 National Institute of Informatics
†2 Clarivate Analytics Co., Ltd.

evaluation organizations use their original subject classifications to classify organizations and persons that fit to domestic evaluation task and internationally compare them based on research and educational output records aggregated in a world common output database, such as Web of Science citation database. For example, the UK government defines units of assessment as subject classifications for the research assessment exercise (RAE) and research excellence framework (REF). Italian evaluation agency for university and research systems, ANVUR made the original category scheme for their evaluations. Australian research evaluation program, excellence in research for Australia (EAR) prepared the original subject classification scheme, fields of research (FoR) and Brazil funding agency, FAPESP created their own subject classification scheme. In educational quality assurance agencies, such as CAPES in Brazil and SCADC in China made their educational subject classifications. All of the above subject classifications are required to be adopted for the Web of Science citation database to analyze their national activities and internationally compare them on the common standard. Along with the present international business needs on the research evaluation, the same requirement emerged from the universities in Japan so that Japanese national funding programs KAKENHI subject classifications would be adopted for the Web of Science citation database.

Adopting subject classifications for bibliographic databases is an extremely hard work. For example, at a time in 2019, the Web of Science citation database in InCites™ – a research output evaluation tool that consists of 58,395,008 article records of 24,688 journals. Even for the articles or the journals as a set of units, assigning subject categories for them are labor-intensive and time consuming. It requires the best way of assigning cost-effectively and efficiently the subject categories on them. Therefore, we aim at the best way and propose an approach to apply a new subject classification scheme for a database. The following sections describes our approach step by step.

2. A Data-driven Approach to Apply a New Subject Classification Scheme

Our approach is data-driven. It is supposed that a subject classification scheme has been originally adopted for a database. Then, it tries to apply a new subject classification scheme for the database by means of a relationship between the two subject classification schemes. The relationship is a correspondence between them, which is induced by data.

2.1 A subject classification model of database

At first, we define a subject classification model of database in order to explain our approach. It is a mathematical formula and a phycological aspect of subject categories embedded in a database.

We suppose that there exists a bibliographic database that represents a set of articles for scientific research. Each article is labeled with at least one category of a subject classification scheme. It means that all articles are classified under the subject classification scheme. The subject classification scheme implies its compact topological space in the database. It states the database structure, which affects analysis by the subject classification scheme.

Definition 1 (a database with a subject classification scheme). A database S is a set of articles a_n. A subject classification scheme C is a set of subject categories c_λ. Articles attributed to a subject category comprise a subset of S, so that 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. Λ is an index set. A subset O_λ depends on the corresponding subject category c_λ. Therefore, we define a map f from a subject classification scheme C to the powerset $\mathfrak{P}(S)$.

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

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

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

Proof. In practical databases, a subject classification scheme C consists of finite elements c_i that are mapped to finite subsets O_i by a map f. Let $\mathfrak{O}$ be a subset of $\mathfrak{P}(S)$ which consists of $\{O_i|i\in I\}$. I is a finite index set. As a basis, let $\mathfrak{O}_0$ be a subset of $\mathfrak{P}(S)$ which consists of $\{\cap_{i\in I}A_i|A_i\in\mathfrak{O}\}$ where the element is S if $I=\emptyset$. Let $\widetilde{\mathfrak{O}}$ be a subset of $\mathfrak{P}(S)$ which consists of $\{\cup_{\lambda\in\Lambda}B_\lambda|B_\lambda\in\mathfrak{O}_0\}$ where the element is $\emptyset$ if $\Lambda=\emptyset$. Λ is a finite or infinite index set. Thus, $\widetilde{\mathfrak{O}}\supset\mathfrak{O}$, $S\in\widetilde{\mathfrak{O}}$, and $\emptyset\in\widetilde{\mathfrak{O}}$. The $\widetilde{\mathfrak{O}}$ is satisfied with the necessary and sufficient conditions to be a topology. In addition to the theorem 1, it implies a compact topological space $(S,\widetilde{\mathfrak{O}})$. When there exists a finite cover in a topological space, we call it as a compact topological space.

2.2 Forming a compact topological space for a new subject classification scheme

Based on the subject classification model of database, we propose an approach to apply a new subject classification scheme for a database.

This time, we suppose the following condition. A subject classification scheme $C^{(1)}$ that consists of subject categories $c_i^{(1)}$ is mapped to a finite cover $\mathfrak{O}^{(1)}=\left\{O_i^{(1)}\middle|i\in I^{(1)}\right\}$ by a map f_1, which implies a compact topological space $(S,\widetilde{\mathfrak{O}}^{(1)})$.

Conventionally, we can take 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. This creates a map f_2 from $C^{(2)}$ to a finite cover $\mathfrak{O}^{(2)}=\left\{O_i^{(2)}\middle|i\in I^{(2)}\right\}$, which implies a compact topological space $(S,\widetilde{\mathfrak{O}}^{(2)})$.

In our approach, we build a correspondence $\Gamma: C^{(2)}\to C^{(1)}$ ($\Gamma=(C^{(2)},C^{(1)};G)$, $G\subset C^{(2)}\times C^{(1)}$), 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 existence of a finite cover.

Then, we create a map

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

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

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

where $S = \bigcup_{i \in I^{(2)}} \bar{O}_i^{(1)}$ to be a finite cover. We get a composite map $g_2 \circ g_1$ from $C^{(2)}$ to a finite cover $\overline{\mathfrak{O}}^{(1)}$, which implies a compact topological space $\left(S, \widetilde{\overline{\mathfrak{O}}}^{(1)}\right)$. Obviously, $\widetilde{\overline{\mathfrak{O}}}^{(1)} \subset \overline{\mathfrak{O}}^{(1)}$.

2.3 Inducing a correspondence between two subject classification schemes by means of a research project database

To decide a correspondence between two subject classification schemes, traditionally experts of the subject classification schemes discuss the relationship structure based on their knowledge and practical experiences.

In our approach, the actor is data scientists who analyze a database where an entity is categorized with the two subject classification schemes and induce the correspondence between them on the analysis.

As evidence data, anything that includes the information indicating the relationship between the two subject classification schemes is useful. One of the candidates is a research project database, which is rather popular among academic databases. So, we follow up our approach to be supposed to adopt a research project database.

2.3.1 By means of a research project database

We define a research project 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.

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

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

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

Then, we create a map

$$h_3: \mathfrak{O}_T^{(2)} \to \mathfrak{O}_{S'}^{(2)} = \left\{ \left. \bar{O}_{S'i}^{(2)} \right| O_{Ti}^{(2)} \in \mathfrak{O}_T^{(2)}, b_j^{(2)} \in O_{Ti}^{(2)},\ O_{S'j}^{(2)} = h_2'\left(b_j^{(2)}\right),\ \bar{O}_{S'i}^{(2)} = \bigcup_j O_{S'j}^{(2)} \right\}$$

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

In this case, we put the following strong suppositions to make it valid. The composite map $h_3 \circ h_1: C^{(2)} \to \mathfrak{O}_{S'}^{(2)}$ represents the classification of articles by the subject classification scheme. And, if two images on S' by a map f_1 and a map $h_3 \circ h_1$ are equivalent, the inverse images of them are of an equivalence relation.

2.3.2 Data-driven approach to induce a correspondence

Now, we have got actual data representing a 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. In a linguistic field, a famous law, named Zipf's law states that a frequency of words obeys a distribution where the word rank n has a frequency proportional to $1/n$. In more general, the same distribution is observed in natural phenomena, named power law, which is denoted as $\ln p(x) = -\alpha \ln x + c$ where α and c are constants [3]. For example, the sizes of city populations, earthquakes, moon craters, solar flares, computer files and wars, the frequency of occurrence of personal names in most cultures, the numbers of papers scientists write, the number of citations received by papers, the number of hits on web pages, the sales of books, music recordings and almost every other branded commodity all follow power law distributions.

Sometimes, when real data is analyzed, in most cases the power law trend holds only for an intermediate range of values; there is a power law breakdown in the distribution tails [4]. This is caused by finite size effects (e.g. insufficient data for good statistics), network dilution, network growth constraints and different underlying dynamical regimes, leading to power law corrections (sometimes referred to as scaling corrections) 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 $f(r) = (A(N+1-r)^b)/r^a$, where r is the rank, N its maximum value, A the normalization constant and (a, b) two fitting exponents.

In our case, elements of finite covers $\mathfrak{O}_{S'}^{(1)}$ and $\mathfrak{O}_{S'}^{(2)}$ represent natural overlapping sets. For an $O^{(2)}(\in \mathfrak{O}_{S'}^{(2)})$, there exist its intersections $O^{(2)} \cap O^{(1)}$ to all $O^{(1)}$ ($\in \mathfrak{O}_{S'}^{(1)}$). Its cardinalities greater than zero, if sorted in rank order, obey the discrete version of the generalized beta distribution since subject categories are finite.

To decide a correspondence between $C^{(1)}$ and $C^{(2)}$, we try to find a subset $\left\{O_i^{(1)} \middle| i \in I_j^{(1)}\right\}$ of $\mathfrak{O}_{S'}^{(1)}$ for an $O_{j\in I^{(2)}}^{(2)}$ to be ideally satisfied that $O_j^{(2)} = \bigcup_{i\in I_j^{(1)}} O_i^{(1)}$. However, in most cases, $O_j^{(2)} \not\supset O_i^{(1)}$ and $O_j^{(2)} \neq \cup O_i^{(1)}$. So, we define the following metrics; (precision)

$$d_p = \frac{\left|\bigcup_{i\in I_j^{(1)}} \left(O_j^{(2)} \cap O_i^{(1)}\right)\right|}{\left|\bigcup_{i\in I_j^{(1)}} O_i^{(1)}\right|}$$

, and (recall)

$$d_r = \frac{\left|\bigcup_{i\in I_j^{(1)}} \left(O_j^{(2)} \cap O_i^{(1)}\right)\right|}{\left|O_j^{(2)}\right|}$$

, and a generalized harmonic mean of precision and recall; (F_β-measure)

$$d_f = \frac{(1+\beta^2)d_p d_r}{\beta^2 d_p + d_r}, \beta > 0.$$

Finally, we decide a threshold of the f-measure to determine which element has a correspondence relation.

3. A Case Study

To verify our approach described above, we adopt it to a practical case. A world leading research output evaluation tool – InCites™ produced by Clarivate Analytics, Co., Ltd., provides bibliometric analysis functions where bibliometrics can be analyzed with domestic subject classification schemes as well as Web of Science subject classification scheme and ESI subject classification scheme. Japanese users are eager to use the subject classification scheme of the biggest Japan's national research grants KAKENHI to analyze their institutional research outputs on the system. The Web of Science citation database holds bibliographic records originally classified with the Web of Science subject classification scheme. The KAKENHI subject classification scheme is a new subject classification scheme to be applied to the Web of Science citation database.

3.1 Induce a correspondence between Web of Science subject categories and KAKENHI subject categories

The followings are the steps we took to induce a correspondence between Web of Science subject categories and KAKENHI subject categories.

3.1.1 Create a contingency table as evidence data

At first, to induce a correspondence between Web of Science subject categories and KAKENHI subject categories, we create a contingency table between them.

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

In this study, we picked up research projects in 2009 whose KAKENHI subject classification scheme consists of a hierarchical structure - 4 categories, 10 areas, 67 disciplines and 284 research fields. The number of projects is 58,952. The number of output publications is 293,753. Of these publications, the number of articles that might be written in English is 173,940.

In KAKEN database, these articles in English are listed in a citation format, which are not yet clear to which Web of Science categories are assigned. So, we identified the same bibliographic records in the Web of Science citation database as of 2009 and 2010 to them by means of a set of record linkage techniques in order to get a set of articles S' that are classified with both KAKENHI subject classification scheme and Web of Science classification scheme as depicted in Figure 1 [5]. The size of the Web of Science citation database we used was 3,925,776, which is classified with 251 subject categories of the Web of Science classification scheme and 22 subject categories of ESI.

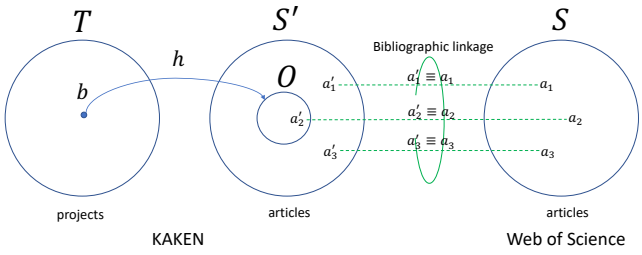


Figure 1. Bibliographic linkage between the KAKEN database and the Web of Science citation database in Venn diagram.

As a result, we got 75,042 pairs of citations, which is 43.1% of 173,940 articles listed in the KAKEN database. The record linkage technique uses i-Linkage as a ranking function and SVM as a classification function to identify the same bibliographic records in KAKEN database and Web of Science citation database. In 10-fold cross validation of 800 samples, the accuracy of the linkage was 95.01. The precision, the recall and the f-measure were 94.92, 95.10 and 94.98, respectively.

Then, we made a contingency table for the two subject classification schemes from the above linkage result, as illustrated in Figure 2. An example in Figure 3 shows a part of the contingency table between the third level 67 KAKENHI subject categories and the 251 Web of Science subject categories.

Of the 75,042 pairs of citations, those which are categorized with the both subject classification schemes are reduced to 59,595

pairs. When the whole counting of the citations to each subject category is applied, we got the sum of 97,175 frequency counts in the contingency table.

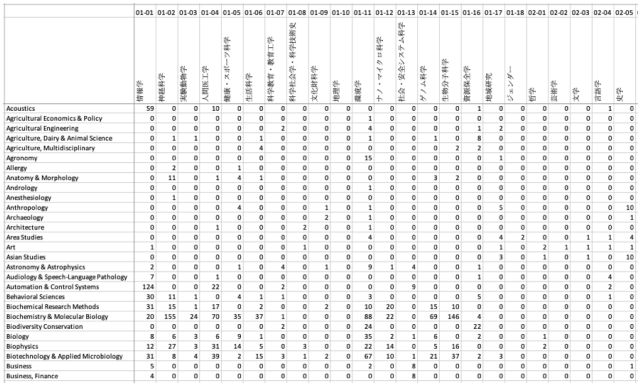
$$f_{ij} = \left| o_i^{(1)} \cap o_j^{(2)} \right|$$

KAKENHI subject categories

Web of Science subject categories

$$\begin{array}{c} \\ o_1^{(1)} \\ \vdots \\ o_m^{(1)} \end{array} \begin{array}{c} \begin{array}{cccccc} o_1^{(2)} & & \cdots & & o_n^{(2)} \end{array} \\ \begin{pmatrix} f_{11} & \cdots & f_{1j} & \cdots & f_{1n} \\ \vdots & \ddots & \vdots & & \vdots \\ f_{i1} & \cdots & f_{ij} & \cdots & f_{in} \\ \vdots & & \vdots & \ddots & \vdots \\ f_{m1} & \cdots & f_{mj} & \cdots & f_{mn} \end{pmatrix} \end{array}$$

Figure 2. A contingency table for the KAKENHI subject categories and the Web of Science subject categories.



	01-01 情報学	01-02 神経科学	01-03 応用数物系	01-04 人間医工学	01-05 機械・ロボット科学	01-06 社会科学	01-07 科学教育・教育工学	01-08 科学社会学・科学技術史	01-09 文化財科学	01-10 地理学	01-11 農業学	01-12 ナノ・マイクロ科学	01-13 社会・安全システム科学	01-14 グリーン科学	01-15 生物分子科学	01-16 環境解析学	01-17 地域研究	01-18 ジェンダー	02-01 哲学	02-02 芸術学	02-03 文学	02-04 言語学	02-05 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	1	0	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	0	3	2	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	5	0	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	0	1	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	2	0	10	20	0	15	10	0	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	0	2	0	8	0	0	0	0	0	0	0	0	1
Business, Finance	4	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0

Figure 3. An example screen of Excel, which shows a part of the contingency table between the third level subject categories of the KAKENHI subject classification scheme and the Web of Science subject classification scheme.

3.1.2 Analysis of the contingency table

To make it clear what happens in the contingency table, we analyzed the distribution among the Web of Science subject categories against a KAKENHI subject category. We observed a good fit of the discrete generalized beta distribution to the rank-ordering distribution in the contingency table.

Figure 4 and 5 shows rank-ordering distributions for the first and the second level of the subject categories of 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), “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 levels, frequencies corresponding to the 251 Web of Science subject categories are sorted in rank order. If the frequency is zero, it is omitted in the distribution. The x axis of the graph represents the rank. The y axis of the graph represents log scale of the frequency count. With these scales, the discrete generalized beta distribution is fitted to data such 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, sets of parameters a and b that affects figures of the distribution vary.

The distributions in the graph can be divided into two types – concentration and dispersal. For the first level of KAKENHI subject categories, the concentration type refers to the graph of “Science and engineering” (l1-03), “Biological sciences” (l1-04). The dispersal type refers to the graph of “Integrated science and innovative science” (l1-01). For the second level, the concentration type refers to the graph of “Humanities” (l2-03), “Chemistry” (l2-06), “Mathematical and physical sciences” (l2-05). The dispersal type refers to the graph of “Comprehensive fields” (l2-01), “New multidisciplinary fields” (l2-02).

For all distributions, good fitness to the discrete generalized beta distribution implies that a set of articles categorized to a KAKENHI subject category naturally overlaps sets of articles categorized to the Web of Science subject categories at any levels. However, the degree of overlapping depends on the target subject categories.

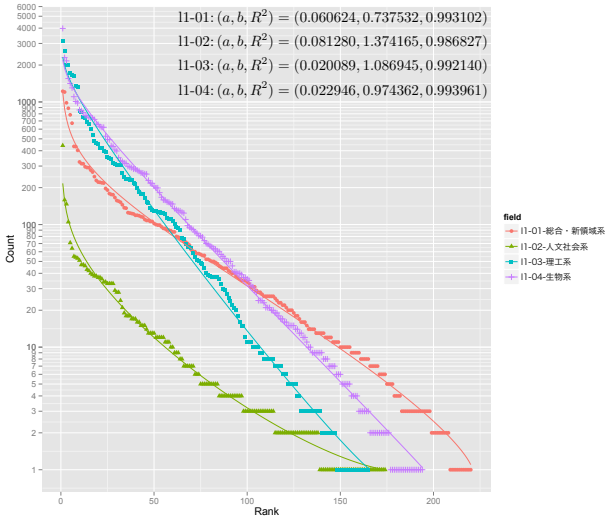


Figure 4. Rank-ordering distributions for the first level subject categories of the KAKENHI subject classification scheme.

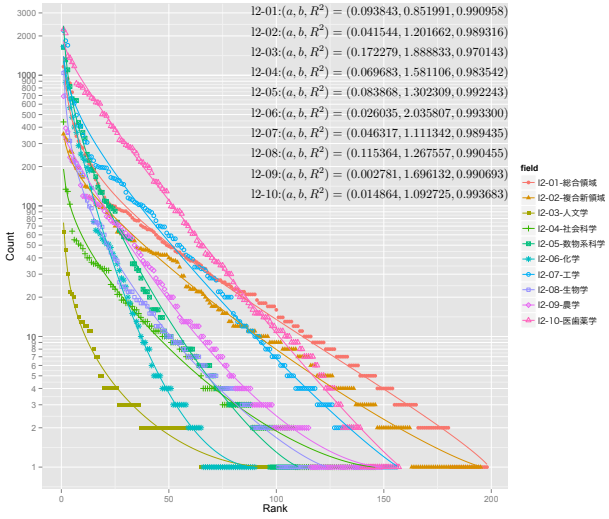


Figure 5. Rank-ordering distributions for the second level subject categories of the KAKENHI subject classification scheme.

3.1.3 Maximizing f-measure

Next, we analyzed how each KAKENHI subject category overlaps the Web of Science subject categories. The aim to induce a correspondence between the KAKENHI subject categories and the Web of Science subject categories encourages us to calculate F_β-measures between them.

Through the contingency table that represents the whole counting of articles, we calculated the following pseudo precision, recall, and F_β-measure based on the original definitions;
(pseudo precision)

$$d'_p = \frac{\sum_i \left| O_j^{(2)} \cap O_i^{(1)} \right|}{\sum_i \left| O_i^{(1)} \right|}$$

, and (pseudo recall)

$$d'_r = \frac{\sum_i \left| O_j^{(2)} \cap O_i^{(1)} \right|}{\left| O_j^{(2)} \right|}$$

, and a generalized harmonic mean of precision and recall; (pseudo F_β-measure)

$$d'_f = \frac{(1+\beta^2) d'_p d'_r}{\beta^2 d'_p + d'_r}, \beta > 0.$$

Appendix lists the maximum pseudo F_1-measure, and the pseudo precision and recall to produce it, and the number of the Web of Science subject categories to cover for the third level 67 disciplines of the KAKENHI subject classification scheme. In this case, the pseudo average precision, recall, maximum F_1-measure were 0.31469, 0.36724, 0.31718, respectively. The number of the Web of Science subject categories to cover a KAKENHI subject category ranges from 1 to 24, which is rather small in comparing to the maximum number 251.

3.1.4 Miscellaneous considerations

In addition to the quantitative analysis above, we set a threshold of article count in the contingency table to ignore relations between the 251 Web of Science subject categories and the 67 disciplines of KAKENHI subject categories. Here, for every Web of Science subject category $O_i^{(1)}$, the number of relations with the KAKENHI subject categories $O_j^{(2)}$ is limited to from 1 to 4 at most. And, in addition, when the recall rate exceeds a half, we stop adding any more relation.

Then, we checked all correspondence between $O_i^{(1)}$ and $O_j^{(2)}$ by means of subject category keywords, especially for the subject categories whose evidence data is very few. The cases are “Arts and Humanities”, “Music”, “Religion”, etc.

Finally, we induced a correspondence between the 10 areas and 67 disciplines of the KAKENHI subject classification scheme and the 251 Web of Science subject categories, which are released in public [6]. There exist 324 relations in-between 10 areas of KAKENHI subject classification scheme and the Web of Science subject categories, and 409 relations in-between 67 disciplines of KAKENHI subject classification scheme and the Web of Science subject categories.

3.2 Classification results on the Web of Science citation database

With the correspondence, a research output evaluation tool InCites™ preprocesses its internal database and provides the functionality of analysis with the KAKENHI subject classification scheme. The followings describe how it provides the analysis function, quantitative statistics of it, and user feedback for the function.

3.2.1 KAKENHI subject categories on InCites™

The tool provides an analytical workbench on the Web of Science citation database. It preprocesses the database to show users target entities such as people, organizations, regions, research areas, journals, books, conference proceedings, funding agencies. Figure 6 is an example screen that shows article counts of Japanese authors by the 67 disciplines of the KAKENHI subject classification scheme. In the figure, bubbles represent top 25 proportional amounts of articles, each of which corresponds to a KAKENHI subject category. The total amount of articles by the Japanese authors is 3,192,449 of the whole 58,395,008 articles published from 1980 to 2018. Of the Japanese authorship, the top KAKENHI subject category at discipline level is clinical internal medicine, which counts 1,096,040. The second and the third are basic medicine and applied chemistry, which count 617,970 and 526,139, respectively.

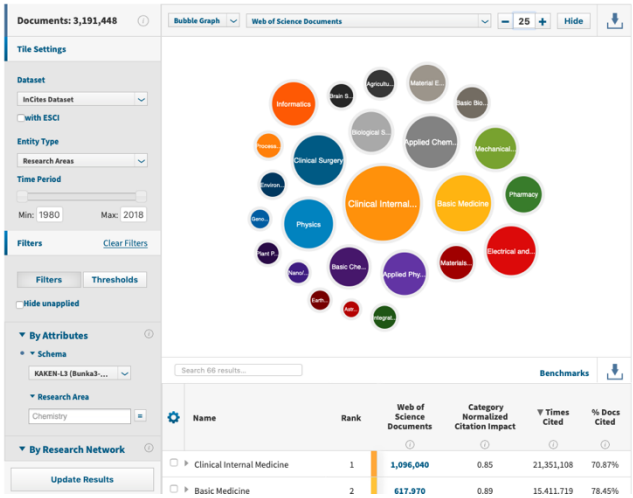


Figure 6. An example screen of InCites™ that shows bubbles representing proportional amounts of articles classified with the KAKENHI subject categories

3.2.2 Article counts by the Web of Science subject categories and the KAKENHI subject categories

For the Japanese authors’ articles, we compared distributions by the subject classification schemes. We illustrated proportions based on the statistics the tool provides by the subject classification schemes in Figure 7, 8, and 9.

Figure 7 shows a top 30 subject distribution of articles by the Web of Science subject classification scheme. From the top, it lists “Engineering, Electrical & Electronic”, “Physics, Applied”, “Biochemistry & Molecular Biology”, “Materials Science, Multidisciplinary”, “Chemistry, Multidisciplinary”, etc. The distribution of the graph gradually declines just like an inverse proportional graph.

Figure 8 shows the subject distribution of the same set of articles by the 10 areas level of the KAKENHI subject classification scheme. From the top, it lists “Medical / Dental / Pharmaceutical”, “Engineering”, “Math / Physics”, “Multi-disciplinary”, “Chemistry”, etc. The number of articles for the subject categories declines linearly rather than inverse proportionally. Figure 9 shows the top-30 subject distribution of the articles by the 67 disciplines level of the KAKENHI subject classification scheme. From the top, it lists “Clinical Internal Medicine”, “Basic Medicine”, “Applied Chemistry”, “Clinical Surgery”, “Electrical and Electric Engineering”, etc. The number of articles declines inverse proportionally. Comparing to the original Web of Science subject categories, this statistical result gives an different impression in that life sciences are stronger among others, although the Web of Science subject classification scheme gives the impression that electrical/electronic engineering and physics are stronger among others.

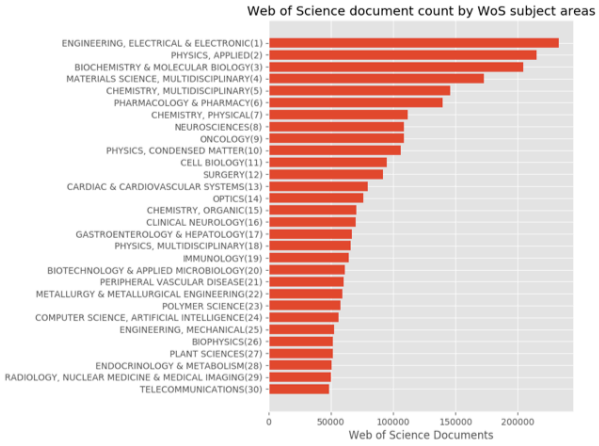


Figure 7. Top-30 subject distribution of the Japanese author’s articles by the Web of Science subject classification scheme

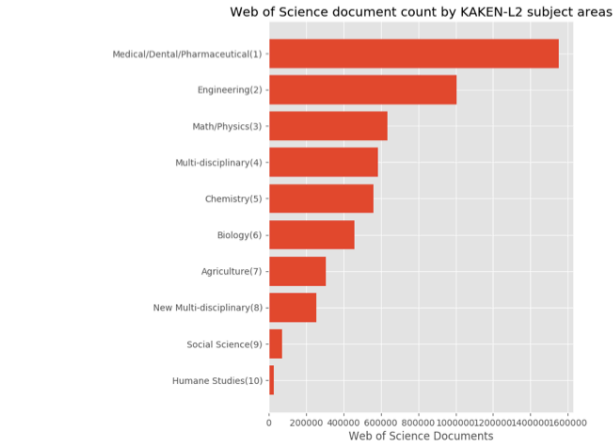


Figure 8. The whole subject distribution of the Japanese author’s articles by the 10 areas level of the KAKENHI subject classification scheme

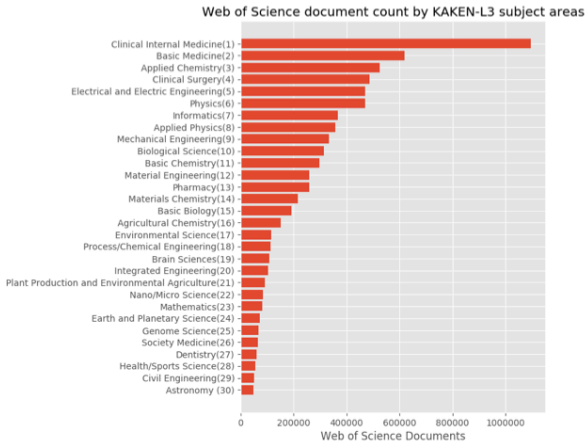


Figure 9. Top-30 subject distribution of the Japanese author’s articles by the 67 disciplines level of the KAKENHI subject classification scheme

3.2.3 User feedback

In response to the KAKENHI subject classification scheme as a new function of InCites™ released in April, 2016, the users in Japan were surveyed by an online questionnaire after a year, in April, 2017.

As a result, 26 institutional users replied the questionnaire, who are mostly research administrators (RAs) and institutional research (IR) staff (Table 1).

Table 1. Users role in their institutions

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

The questionnaire consists of 18 questions related to the subject classification schemes implemented in InCites™ and the attributes of users. An open answering question was set in its end.

To know about the degree of expertise of the users, Q13 and Q3 were prepared. Q13 asked how often users use the InCites™. Q3 asked how much the users know about the KAKENHI subject classification scheme. Most of the users periodically use the tool in their work, and know well about the KAKENHI subject classification scheme.

About the validity of the KAKENHI subject classification scheme, Q7 and Q11 were set. Q7 asked which level of the hierarchy of the KAKENHI subject classification scheme is needed. Q11 asked whether the users feel comfortable with their analysis results by the KAKENHI subject classification scheme. As a result, it is revealed that the users think they need both levels of hierarchy, and they almost feel comfortable with their analysis results by the KAKENHI subject classification scheme comparing to their experience in KAKENHI funding related jobs.

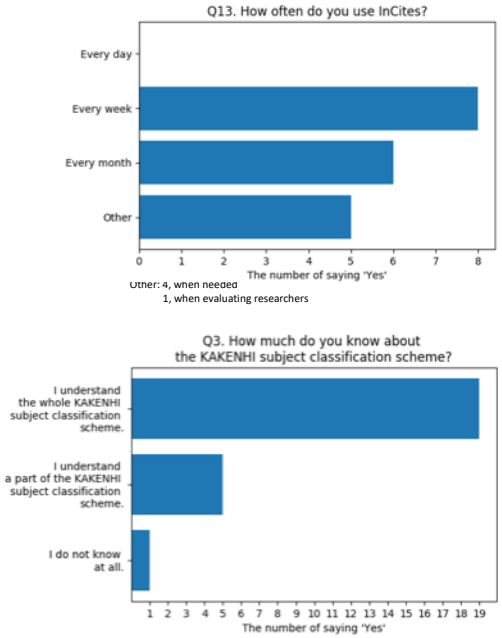


Figure 10. Questions and answering results for the user's degree of expertise for InCites™ and the KAKENHI subject classification scheme

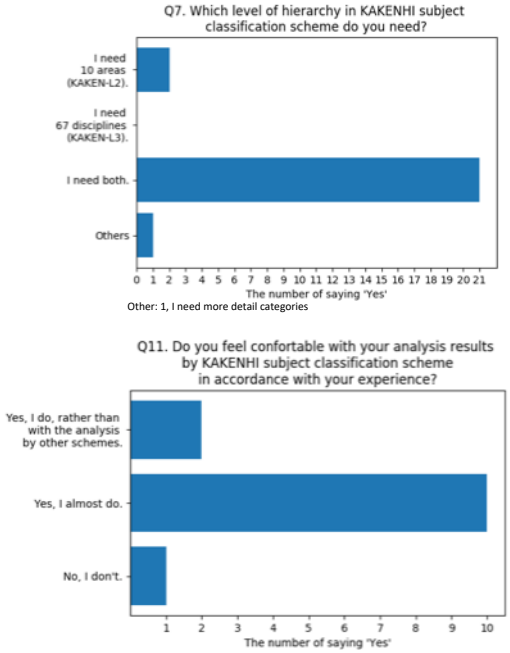


Figure 11. Questions and answering results for the validity of the KAKENHI subject classification scheme

In the free answering question asking for the additional comment on the KAKENHI subject classification scheme as a new feature, many users insisted that they need it. For further demands, they expressed that they need the same subject categories for the other services, and want it to be updated, and more precise one in such the following comments;

- *"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 ES (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."*

3.3 Discussion

When we look at the theory of our approach, i.e., inducing a correspondence between two subject classification schemes, we recognize that it has an inherent limitation. The embedding subject classification scheme inevitably depends on the original classification scheme. The topological space of the former is a subset of the topological space of the latter. Unfortunately, we observed that in natural correlations between subject categories of two subject classification schemes, each subject category of one scheme partly overlaps several subject categories of the other scheme. There is no inclusion relationship between them. Thus, it implies that correspondence relations must be probabilistic.

And, we have set strong suppositions on relations among research projects and journal articles in the research project database, in that they have similarities on subject. But in fact, they have similarities and differences on subject. On the side of the similarity, we admitted the following procedure. A grants database describes that research projects produce outputs, i.e., research articles. We focused on the subject classification scheme for the research projects and its relationship to a set of research articles. Research articles are classified with another subject classification scheme. Then, we compared those two subject classification schemes through its relationship. On the side of the difference, we have another story. Projects precede articles. There is a time lag of project starting and article outputs. This makes a subject divergence or drift between them. And, projects tend to indicate the central concept with essential keywords. This allows a subject diversification of articles.

Nevertheless, the users of InCites™ accepted the subject classification results. We imagine some reasons as follows. Users might focus on comparative analysis of bibliometrics by the subject categories, and not care about specific case of articles. They might need rough quality of metrics at the evaluation stage. Metrics are central limits of quantitative attributes of a set of entities, which is the main indicator to be checked for the research evaluation.

Another advantage is that our approach is extremely cost effective. At a time of 2019, the number of the Web of Science documents stored in InCites™ is 58,395,008, whose journal titles amount to 24,688. So far, the possible targets to assign subject categories are the Web of Science documents and journal titles. Journal titles include a set of documents. Assigning subject categories to journal titles means consequently assigning them to documents. In production, the Web of Science subject categories are assigned to mainly journal titles and exceptionally documents in multidisciplinary journals. In our approach, we induced a

correspondence between the Web of Science subject classification scheme and the KAKENHI subject classification scheme by means of the KAKEN database. For the 251 Web of Science subject categories and 67 disciplines of the KAKENHI subject categories, the maximum relations in the correspondence count up to 16,817 (251×67). As for the 10 areas of KAKENHI subject categories, the maximum relations count up to 2,510 (251×10). The number to check relations in our approach is overwhelmingly smaller than that of the original subject category assignment approach.

The evidence data is the contingency table whose sum of the frequency counts is 97,175. In fact, this number is not sufficient for an automatic decision making because when we checked the correspondence between both subject classification schemes, there existed apparently lacks of relations between them although the relations ought to exist in the literary meanings. Manual handling was needed for some subject categories. If the data size would become large enough, we could predict the correspondence only by the data.

4. Conclusions and Future Work

We proposed an approach to apply a new subject classification scheme for a bibliographic database that is already classified by a subject classification scheme. In this paper, we defined a subject classification model of database that consists of a topological space. Then, we showed our approach based on the model, where the step is to form a compact topological space for a new subject classification scheme. To form the space, it utilizes a correspondence between two subject classification schemes by a research project database as data.

We applied the approach to a practical example, i.e. InCites™ - a world proprietary benchmarking tool for research evaluation based on the Web of Science citation database so as to add the subject classification scheme of Japan's national biggest grants KAKENHI. The Web of Science subject classification scheme consists of 251 subject categories, and the KAKENHI subject classification scheme consists of a hierarchy – 4 categories, 10 areas, 67 disciplines, and 284 research fields. We created two correspondences between 10 areas / 67 disciplines of the KAKENHI subject categories and the 251 Web of Science subject categories. By means of the KAKEN database that keep records of research project descriptions and achievement lists of KAKENHI and a set of record linkage techniques, i.e., i-Linkage and SVM, 59,595 pairs of articles classified with the both subject classification schemes are extracted. Then, we analyzed the pairs of articles so as to induce the correspondences between the 10 areas / 67 disciplines of KAKENHI subject categories and 251 Web of Science subject categories based on our approach. The InCites™ give a function of analysis with the KAKENHI subject classification scheme. By a user survey, it is revealed that the users accepted the feature on the whole.

As for future work, there are several aspects for demanding the quality of database and embedding subject classification schemes by means of effective and efficient automatic procedures. As ever, metadata is a good tool for information management and analysis. It describes entities at an abstract level, incorporates necessary context, and equips analytical viewpoints. Originally, metadata is described by information professionals. In present data age, it will be handled on the basis of external data and artificial intelligence. Our approach become robust by large amount of data. In an alternative way, it is promising to directly look into content and extract knowledge for the same purposes on metadata.

Reference

[1] Hjørland, B. (2008). What is Knowledge Organization (KO)? Knowledge Organization, 35(2–3), 86–101. http://doi.org/10.5771/0943-7444-2008-2-3-86

[2] Gómez, I., Bordons, M., Fernández, M. T., Méndez, A. (1996). Coping with the problem of subject classification diversity. Scientometrics, 35(2), 223–235. http://doi.org/10.1007/BF02018480

[3] Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. Contemporary Physics, 46(5), 323–351. http://doi.org/10.1080/00107510500052444

[4] Martínez-Mekler, G., Alvarez Martínez, R., Beltrán del Río, M., Mansilla, R., Miramontes, P., Cocho, G. (2009). Universality of rank-ordering distributions in the arts and sciences. PloS One, 4(3), e4791. http://doi.org/10.1371/journal.pone.0004791

[5] Kurakawa, K., Sun, Y., 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. https://www.nii.ac.jp/TechReports/public_html/14-002J.html

[6] Clarivate Analytics, "KAKEN Category Scheme - InCites Help", http://help.prod-incites.com/inCites2Live/filterValuesGroup/researchAreaSchema/kaken.html, (accessed 2019-02-21).

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 National Institute of Informatics, Cyber Science Infrastructure Development Department, Scholarly and Academic Information Division, 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.

Appendix

Table 2. Maximum pseudo F_1-measure for the third level 67 disciplines of the KAKENHI subject categories against the 251 Web of Science subject categories

10 areas – 67 disciplines seq. no.	KAKENHI subject category	Translation	# of WoS subject categories to cover	Pseudo precision	Pseudo recall	Max pseudo F1 measure
(01-01)	情報学	Informatics	17	0.57582	0.62589	0.59981
(01-02)	神経科学	Brain sciences	1	0.21829	0.36497	0.27318
(01-03)	実験動物学	Laboratory animal science	1	0.05863	0.07438	0.06557
(01-04)	人間医工学	Human informatics	8	0.22199	0.21253	0.21716
(01-05)	健康・スポーツ科学	Health / sports science	5	0.18095	0.29028	0.22293
(01-06)	生活科学	Human life science	4	0.23905	0.28051	0.25813
(01-07)	科学教育・教育工学	Science education /educational technology	2	0.37736	0.10309	0.16194
(01-08)	科学社会学・科学技術史	Sociology / history of science and technology	6	0.11111	0.16279	0.13208
(01-09)	文化財科学	Cultural assets study	1	0.2	0.03636	0.06154
(01-10)	地理学	Geography	4	0.11719	0.2027	0.14851
(01-11)	環境学	Environmental science	14	0.26227	0.3853	0.3121
(01-12)	ナノ・マイクロ科学	Nano / micro science	4	0.10326	0.31317	0.15531
(01-13)	社会・安全システム科学	Social / safety system science	14	0.18656	0.21429	0.19946
(01-14)	ゲノム科学	Genome science	3	0.04047	0.20305	0.06748
(01-15)	生物分子科学	Biomedical engineering	2	0.11913	0.32457	0.17429
(01-16)	資源保全学	Culture assets and museology	3	0.18116	0.14535	0.16129
(01-17)	地域研究	Area studies	7	0.16429	0.27059	0.20444
(01-18)	ジェンダー	Gender	3	0.23077	0.11111	0.15
(02-01)	哲学	Philosophy	4	0.4359	0.28333	0.34343
(02-02)	芸術学	Art studies	1	0.09091	0.11111	0.1
(02-03)	文学	Literature	10	0.7	0.68293	0.69136
(02-04)	言語学	Linguistics	3	0.70504	0.41004	0.51852
(02-05)	史学	History	6	0.41176	0.34146	0.37333
(02-06)	人文地理学	Human geography	3	0.175	0.5	0.25926
(02-07)	文化人類学	Cultural anthropology	3	0.05634	0.10526	0.07339
(02-08)	法学	Law	3	0.38462	0.12195	0.18519
(02-09)	政治学	Politics	2	0.40909	0.45763	0.432
(02-10)	経済学	Economics	12	0.6917	0.62198	0.65499
(02-11)	経営学	Management	5	0.29412	0.38462	0.33333
(02-12)	社会学	Sociology	8	0.17606	0.27778	0.21552
(02-13)	心理学	Psychology	14	0.4878	0.47859	0.48315
(02-14)	教育学	Education	9	0.24375	0.25828	0.2508
(03-01)	数学	Mathematics	4	0.73424	0.79181	0.76194
(03-02)	天文学	Astronomy	1	0.5052	0.86965	0.63912
(03-03)	物理学	Physics	6	0.49831	0.65128	0.56462
(03-04)	地球惑星科学	Earth and planetary science	7	0.6186	0.66222	0.63967
(03-05)	プラズマ科学	Plasma science	1	0.23261	0.19094	0.20973
(03-06)	基礎化学	Basic chemistry	7	0.22929	0.80065	0.35649
(03-07)	複合化学	Applied chemistry	6	0.28307	0.52645	0.36817
(03-08)	材料化学	Materials chemistry	7	0.1571	0.34801	0.21647
(03-09)	応用物理学・工学基礎	Applied physics	5	0.17011	0.39374	0.23758
(03-10)	機械工学	Mechanical engineering	11	0.43053	0.38804	0.40818
(03-11)	電気電子工学	Electrical and electric engineering	10	0.33758	0.66933	0.4488
(03-12)	土木工学	Civil engineering	8	0.37069	0.48383	0.41977
(03-13)	建築学	Architecture and building engineering	3	0.28571	0.50588	0.36518
(03-14)	材料工学	Material engineering	6	0.34794	0.52269	0.41778
(03-15)	プロセス工学	Process / chemical engineering	4	0.14529	0.30553	0.19694
(03-16)	総合工学	Integrated engineering	8	0.25637	0.30922	0.28032
(04-01)	基礎生物学	Basic biology	7	0.375	0.39992	0.38706
(04-02)	生物科学	Biological science	4	0.16679	0.58193	0.25927
(04-03)	人類学	Anthropology	3	0.31504	0.44	0.36718
(04-04)	農学	Plant production and environmental agriculture	4	0.30676	0.44939	0.36462
(04-05)	農芸化学	Agricultural chemistry	6	0.22042	0.38632	0.28069
(04-06)	林学	Forest and forest products science	5	0.40751	0.25224	0.3116
(04-07)	水産学	Applied aquatic science	2	0.4185	0.32702	0.36715
(04-08)	農業経済学	Agricultural science in society and economy	2	0.33333	0.09677	0.15
(04-09)	農業工学	Agro-engineering	4	0.15686	0.25926	0.19546
(04-10)	畜産学・獣医学	Animal life science	4	0.51054	0.38655	0.43998
(04-11)	境界農学	Boundary agriculture	4	0.2346	0.14787	0.18141
(04-12)	薬学	Pharmacy	4	0.29417	0.3694	0.32752
(04-13)	基礎医学	Basic medicine	16	0.21266	0.55141	0.30695
(04-14)	境界医学	Boundary medicine	12	0.16156	0.11176	0.13213
(04-15)	社会医学	Society medicine	8	0.28153	0.26197	0.2714
(04-16)	内科系臨床医学	Clinical internal medicine	24	0.44074	0.61743	0.51433
(04-17)	外科系臨床医学	Clinical surgery	20	0.41795	0.468	0.44156
(04-18)	歯学	Denticity	3	0.64007	0.27983	0.38941
(04-19)	看護学	Nursing	2	0.73684	0.44304	0.55336
Average				**0.31469**	**0.36724**	**0.31718**