

A search strategy for publications in interdisciplinary research

Wenjing Xiong; Ping Zhou

Recommended citation:

Xiong, Wenjing; Zhou, Ping (2023). "A search strategy for publications in interdisciplinary research". *Profesional de la información*, v. 32, n. 5, e320522.

<https://doi.org/10.3145/epi.2023.sep.22>

Article received on May 19th 2023
Approved on September 27th 2023



Wenjing Xiong

<https://orcid.org/0000-0002-8837-5178>

Zhejiang University
School of Public Affairs
Dept. of Information Resources Management
No. 866 Yuhangtang Road
310059 Hangzhou, China
wenjing_xiong@zju.edu.cn



Ping Zhou ✉

<https://orcid.org/0000-0001-5471-5261>

Zhejiang University
School of Public Affairs
Dept. of Information Resources Management
No. 866 Yuhangtang Road
310059 Hangzhou, China
pingzhou@zju.edu.cn

Abstract

To retrieve the right collection of publications in interdisciplinary research, we have developed a search strategy with four progressive steps and take the area of public affairs (PA) as a case study. A set of seed publications in PA is first identified, followed by the construction of a pool set of publications with wider coverage for refinement in the next step, which is critical and in which an expanded set of publications is established on the basis of the references and text semantic information, thus generating two respective subsets. One of these subsets is obtained on the basis of the number of references shared between each publication pair between the seed set and the pool set. To optimize the results, we construct two models, viz. a support vector machine (SVM) and a fully connected neural network (FCNN), and find that the FCNN model outperforms the SVM model. The second subset of publications are collected by selecting the publications with high topic similarity to the seed publications collected in the first step. The final step is to integrate the seed publications with the expanded publications collected in steps 1 and 3. The results show that PA research involves an extremely wide range of disciplines ($n = 45$), among which public administration, environmental sciences, economics, management, and health policy and services, among others, play the most significant roles.

Keywords

Search strategies; Information retrieval; Public affairs; Interdisciplinary research; Field delineation; Bibliometric analysis; Publications; Scientific research.

Funding

This study received financial support from the *National Natural Science Foundation of China* (NSFC, no. 71843012).

1. Introduction

Scientific research involving multiple disciplines/fields plays a significant role in the development of science and the social sciences, thus leading to numerous studies on interdisciplinary research from various perspectives (Lovelock; Margulis, 1974; Wagner *et al.*, 2011; Wang *et al.*, 2015; Glänzel; Debackere, 2021; Ledford, 2015). How to define the disciplinary attribution of interdisciplinary research remains a challenge because the involved knowledge is from two or more disciplines (Klein *et al.*, 1997; Repko *et al.*, 2007; OECD, 1972; Rotolo *et al.*, 2015) that may vary depending on the specific research topic (Glänzel *et al.*, 2016). It thus becomes difficult to retrieve interdisciplinary publications on the basis of existing classification schemes such as *Web of Science* (WoS) subject categories. Working at the level of scientific journals, the WoS scheme consists of approximately 250 research areas equivalent to subject categories. A multidisciplinary journal may be assigned to one or more research areas (Leydesdorff; Ràfols, 2009). To improve the WoS journal-based system, some science and technology organizations have developed their own, journal-level classification systems (Archambault *et al.*, 2011; Glänzel; Schubert, 2003; Boyack; Klavans, 2010). Such classification schemes, however, cannot be used directly to define interdisciplinary publications.

Classification schemes at the publication level may better reflect research subjects and have been explored at a relatively small scale in the early period (e.g., Griffith *et al.*, 1974; Small; Griffith, 1974; Small; Sweeney, 1985), until 2010 when a



large-scale publication-level classification scheme was proposed (Boyack; Klavans, 2010). On the basis of direct citation relationships between publications, the Leiden classification system covers scientific fields (Waltman; Van-Eck, 2012) and is composed of three hierarchies, with 20 research areas at the first level, 672 research areas at the second level, and 22,412 research areas at the third level. Working at publication level instead of journal level, the Leiden classification system can classify specific scientific areas in more detail and thus match the current structure of scientific research more closely. Using the Leiden methodology, interdisciplinary publications can be harvested by following simple steps. The requirements in terms of computing time and memory usage, however, are too demanding for individual researchers using a standard desktop computer. In addition, the Leiden methodology does not consider indirect citation relations, which may lead to an incomplete collection of interdisciplinary publications.

To collect interdisciplinary publications on specific topics such as innovation systems, big data, or public affairs, five methods can be considered. The first is to identify historical core contributions and community members in a specific area (Fagerberg *et al.*, 2012). With an implicit assumption that the historical core literature remains stable over time, such an approach is not able to reflect the dynamic changes in a specific field. The second relies on a set of keywords and has wider applications (e.g., Huang *et al.*, 2011; 2015; 2019; Ruiz-Navas; Miyazaki, 2018; Liu *et al.*, 2021). Variations exist in keyword approaches; For example, some researchers regard keywords as a core set that defines a specific field (Huang *et al.*, 2011), which requires strong knowledge of the field and the consideration of emerging terms. By making full use of the feature of semi-automatic iteration of keyword variants, some researchers first apply a keyword retrieval strategy to identify a core set of publications and then supplement or modify this core set by using high-frequency words (e.g., Liu *et al.*, 2021; Suominen *et al.*, 2016; Shapira *et al.*, 2017), consulting experts in related fields (Huang *et al.*, 2015), or using synonyms and subordinate words through a retrieval formula (Ruiz-Navas; Miyazaki, 2018). The third is to expand related publications through direct citation relations among publications or journals (e.g., Waltman; Van-Eck, 2012; Muñoz-Écija *et al.*, 2019; Bassecouard *et al.*, 2007), which requires high computational and memory capacity. To avoid such capacity demands, researchers start with a core set of publications and then collect publications linked by citing and cited relationships (Wang *et al.*, 2019). However, this may result in a large number of redundant documents with low relevance to the research topic being included in the core document set. In addition, starting from a core set of publications, the fourth takes advantage of co-citation relations and bibliographic coupling (Kostoff *et al.*, 2006; Zitt; Bassecouard, 2006; Alencar *et al.*, 2007; Soós; Kiss, 2020; Zhou *et al.*, 2019). This approach, however, is sensitive to the coverage of the starting corpus, in addition to the time lag of citations (Mogoutov; Kahane, 2007).

As a comprehensive retrieval strategy with multiple rounds of iterations (Glänzel, 2015), the fifth includes two critical tasks and is more accurate than the previous four. Step 1 is to determine a core set of documents (Zitt; Bassecouard, 2006) that can represent the subject in question well, and step 2 is to extend this core set with relevant documents on the basis of thematic similarity through different approaches. To ensure that the documents collected in the second step are relevant to the subject area, a bibliometric-added retrieval method was proposed, combining complex structures instead of individual search terms. All types of search fields, including keywords, terms, subject headings, journal titles, citations and references, and even organization addresses and author names/identifiers can be incorporated into this retrieval strategy. Taking the approach proposed by Rakas and Hain (2019), for example, they first collected a seed set of publications successively through keyword retrieval and highly cited publications, and then collected a set of “relevant documents” on the basis of overlapping bibliography (i.e., two articles with a high degree of overlap with a larger number of identical references). The two sets of publications are integrated as a final corpus for subsequent analysis. Duplicates and the publications that were not cited at least once are excluded. However, defects can exist in the process of publication collection because documents that are important but that have not yet attracted community attention may be missed. Moreover, extracting 500 publications with the highest bibliographical overlap with each seed publication may result in the inclusion of publications that are irrelevant or less relevant to the topic under consideration.

The diversity and variety of interdisciplinary research (Leydesdorff *et al.*, 2019) result in difficulty when it comes to the classification and retrieval of interdisciplinary publications. Most existing search strategies have pros and cons. The current paper proposes a general search strategy that can be applied to any interdisciplinary subject or field, and then subsequently applies this strategy to the discipline of public affairs (PA), a typical interdisciplinary area.

2. A general search strategy

The disciplines or fields involved in interdisciplinary research may vary significantly in terms of their number and type, resulting in various search strategies, such those introduced above. In the case where no single approach is widely accepted, it becomes necessary to propose a framework that is applicable to any interdisciplinary research field. Below, we first introduce such a general search strategy with four steps, and then apply the strategy to the publication collection of public affairs. Compared with other interdisciplinary subjects or fields, the interdisciplinary situation of public affairs is more complex in terms of both the variety and diversity of the disciplinary/field. A search strategy applicable to PA can thus be easily adjusted to other interdisciplinary subjects/fields with less diversity or variety.

A strategy should consider both recall and precision. According to Bradford’s law and by summarizing the pros and cons of existing methods, we propose a framework with four steps: Emphasizing precision, the first step is to construct a seed

set of publications. The second and third steps take recall into consideration but with different focuses: step 2 aims to construct a pool set of publications so as to enlarge the coverage, whereas step 3 aims to construct an expanded set of publications by extracting publications from the pool set that are similar to the target discipline or field. In the fourth step, a final corpus of publications is obtained by integrating the seed publications and expanded publications.

The seed publications collected in the first step should be the most representative of the target discipline or field, to ensure the relevance of the expanded publications collected in next steps. The methods to be chosen are dependent on the degree of interdisciplinarity, for instance, with or without a clear core field/discipline. For those with a clear core discipline or field (e.g., public affairs or digital finance), a list of representative journals –core journals according to Bradford’s law, subject category, or other combined methods– can be applied; for those without a definite core field or discipline (e.g., digital governance or innovation systems), a retrieval method using a set of topic words can be applied. To ensure the representativeness of the obtained collection, other bibliometric methods (e.g., top-high citations) can be applied.

To establish a pool set of publications with wider coverage, a well-accepted subject classification system with the targeted interdisciplinary category can be used. In addition to that of the *WoS*, national and organizational subject classification systems may also work. For example, the classification system of the *National Natural Science Foundation of China (NSFC)* is used in the current study. All the mentioned classification systems are based on journals.

Nevertheless, the problem with such types of classification is obvious, because not all publications in journals of the same interdisciplinary subject category are necessarily similar. Thus, publications with similarity in the pool set should be identified and extracted, which is the task of step 3: constructing an expanded set by assessing the subject similarity of publications in the pool with each of the seeds. An assessment of either reference similarity or topic similarity of the publications in the two sets (i.e., the pool set and the seed set) can serve this purpose. To assess reference similarity, publications in the two sets are paired to calculate dynamic thresholds for the number of common references by adopting models such as a support vector machine (SVM) or a fully connected neural network (FCNN). The operational results will indicate which model should be used. On the basis of the dynamic threshold, similar publication pairs can be identified. To evaluate topic similarity, a text semantic similarity algorithm (Jaccard, Word2vec, and TF-IDF) may be an option.

3. The case of public affairs

After more than 100 years of development, the discipline of public affairs has continuously absorbed knowledge from numerous disciplines/fields, such as political science, management, law, psychology, sociology, economics, and information science (Benz, 2005; Rodgers; Rodgers, 2000; Fleisher; Blair, 1999; Harris; Moss, 2001a; Harris; Fleisher, 2005), with the results of PA being regarded as a “borrowing discipline” (Stallings, 1986). With economic development and social progress, PA has been facing increasingly new problems and challenges (Raadschelders, 2011). Many sophisticated policy issues simply cannot be addressed within the narrow boundaries of traditional disciplines. Absorbing knowledge from a broader scope of subjects, including the natural sciences and engineering, has accompanied the development of PA as a discipline, and thus further increases its interdisciplinarity (Kettl; Milward, 1996). Some “distant” subjects, such as computer science, mathematics, and statistics, have become an inexhaustible resource for PA (Savage, 1974). Undoubtedly, the extensive assimilation of knowledge contributes significantly to the development of PA; in the meantime, however, there is difficulty in delineating the field composition of the discipline (Yang, 2019; Steckmest, 1982; Harris; Moss, 2001b). Failing to clarify this field/disciplinary attribution may limit researchers’ reference scope and thus affect the development of the discipline, weakening scholars’ sense of field belonging and even pushing PA into a crisis of discipline extinction over time (Ostrom, 1974; White et al., 1996; Seibel, 1996; Denhardt, 2004). In addition, the field variety within PA is significant owing to the broad inclusion of fields or disciplines such as public administration, environment protection, social security, digital governance, etc. Different public affairs universities may vary in field coverage, which results in enormous challenges to benchmarking practices (e.g., university ranking and performance evaluation) because of the dependence of the field on research evaluation (Hicks et al., 2015; Gómez-Núñez et al., 2014).

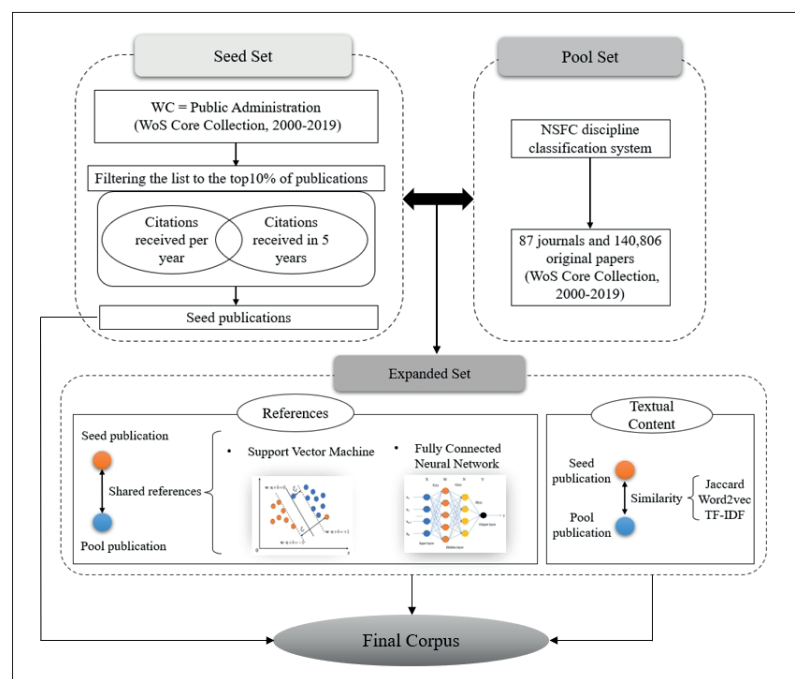


Figure 1. Search framework for publications in public affairs

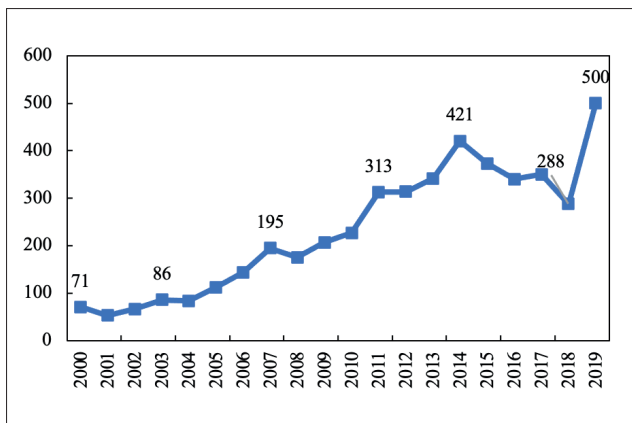


Figure 2. Annual development of the seed publications

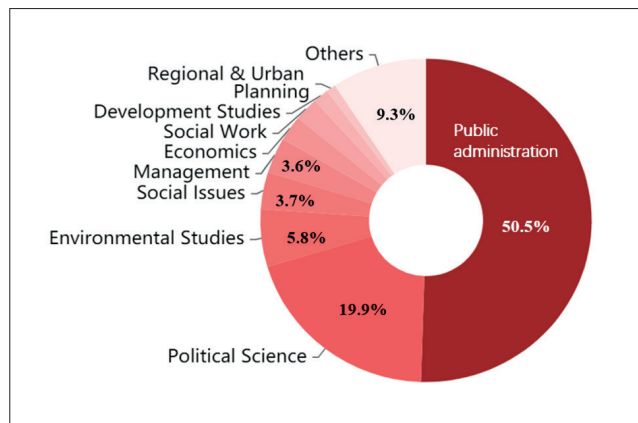


Figure 3. Discipline distribution of the seed publications

To collect publications in public affairs, we follow the general search strategy proposed above. The detailed framework is illustrated in Figure 1.

4. Step 1. Defining a seed set of publications

Public affairs involves various relationships surrounding governments (Fleisher, 2001; Harris; Fleisher, 2005; Lerbinger, 2006), with public administration as the core. Thus, public administration can be regarded as the core field of public affairs. Articles indexed in the *WoS* and published in the years 2000-2019 in the public administration subject category are first collected. Among the obtained publications, only those with top 10% citations in 5 years or per year are considered as the seed set of publications of public affairs. After removing duplicates, 4,664 seed publications are confirmed. The historical development and discipline distribution of the seed publications are shown in Figures 2 and 3. Over 10 years, the seed publications of PA have increased about seven fold. PA itself is multidisciplinary, involving more than 10 disciplines, and more than half of the seeds involve research in the public administration field. Political science is also one of the main fields of PA.

5. Step 2. Constructing a pool set of publications

Defining the seed set of publications by using highly cited articles may result in a lack of knowledge regarding other important contributions. It is necessary to construct a pool set of publications with wider coverage by searching the journals related to public affairs. The list of 91 journals in the current study is provided by the *Development Strategy for the Discipline of Public Affairs* (Xue et al., 2017), an output of a key project of the *National Natural Science Foundation of China*. On the basis of the journal list that is widely accepted by universities, as well as the practice of Yu (2019), four journals (energy, applied energy, agricultural water management, and energy economics) were eliminated owing to the keyword clustering analysis topics of the publications in the four journals mainly being related to energy, resource, and environmental technology, with little relevance to public affairs. Furthermore, the publication sizes of these four journals are unstable and have increased significantly from several hundred to more than 4,000 since 2010, which may affect the statistical results of PA research. The other 87 journals are indexed in the *Web of Science* (*WoS*), and include a total of 140,806 articles published in the period 2000-2019. In other words, the pool set comprises 140,806 articles. The wide disciplinary coverage of publications in the pool set is presented in Table 1. The public administration field does not take first position anymore; instead, environment-related fields and economics take the lead. The purpose of enlarging the disciplinary coverage has been reached in a reasonable way.

Table 1. Distribution of top 20 disciplines in the pool

Discipline	N	Ratio
Environmental sciences	41,159	13.5%
Environmental studies	34,339	11.2%
Economics	26,401	8.6%
Public, environmental, and occupational health	16,023	5.2%
Public administration	13,665	4.5%
Health policy and services	13,208	4.3%
Healthcare sciences and services	12,559	4.1%
Ecology	10,297	3.4%
Energy and fuels	9,781	3.2%
Information science and library science	9,675	3.2%
Management	9,605	3.1%
Education and educational research	8,504	2.8%
Business	6,205	2.0%
Ergonomics	6,106	2.0%
Social sciences, interdisciplinary	5,872	1.9%
Transportation	5,798	1.9%
Geography	5,082	1.7%
Political science	4,969	1.6%
Computer science, interdisciplinary applications	4,926	1.6%
Agriculture, multidisciplinary	4,739	1.6%

6. Step 3. Establishing an expanded set of publications

The wider coverage of the pool set may include less or irrelevant documents. To ensure the publications are indeed relevant to public affairs, we adopt two methods to establishing an expanded set of publications, namely measuring the number of shared references and measuring the topic similarity between the seed publications and those in the pool set. Measuring shared references is more complicated and takes longer to illustrate.

Measurement of shared references. The references of the seed publications and the pool publications are compared in pairs. The number of shared references is used to determine whether two publications are similar. With regard to the appropriate threshold for the shared number of references, variation exists among different studies. The detailed operation is as follows: we first randomly select a sample of 500 publications from the seed set, and then select 500 publications from the pool set and pair them with each initial “seed.” A dataset of 500 publication pairs (one-to-one matching) is obtained. Second, on the basis of the similarity of the title, abstract, keywords, and other information, we manually label whether each pair of publications is similar (1 for similar and 0 for not similar pairs). In the meantime, the number of references of each pair and the number of shared references are recorded. **Rakas and Hain (2019)** define the top 500 publications with a high overlap rate with the references in the pool set as the expanded publications of each seed publication. The disadvantages of this method are obvious: a lack of knowledge of citation disparities among document types, the time length of publication, and research topics. Such disadvantages would become severe when the number of shared references is small and where the number of references between publications varies greatly. The similarity between the target publications (a seed publication and a pool publication) in cases A and B in Table 2 would be considered equal on the basis of an absolute threshold, although it is clear that significant variation exists.

Table 2. Determination of similar publications based on an absolute threshold

Case	Seed R	Pool R	Shared R
A	34	294	1
B	4	7	1

Note: Calculated according to the method of **Rakas and Hain (2019)**.

Table 3. Determination of similar publications based on a relative threshold

Case	Seed R	Pool R	Shared R: 20%*Min (Num- -seedR, Num-poolR)
C	100	100	20
D	150	140	28
E	100	10	2

Could a relative threshold be a better solution? We use three cases, presented in Table 3, for further understanding. When the difference in the number of references between the two compared papers is small, a relative threshold can be set to roughly determine the similarity on the basis of a smaller number of references [i.e., similarity = 20% × Min (Num-Seed R, Num-Pool R)]. Thus, in case D, the number of references shared between a seed publication and a pool publication is 28. When the difference is large, however, a relative threshold does not work, for instance, case E in Table 3.

Given that neither absolute nor relative thresholds can be used to reasonably determine publication similarity, this study tries to construct high-dimensional models for the determination of publication similarity. Two models, a support vector machine and a fully connected neural network, are constructed and compared to select the one that is optimal. The result of each model is compared with that of the method proposed by **Rakas and Hain (2019)**.

On the basis of sample records of 500 publications selected randomly and indexed manually, a formula is developed to simulate the relationship reflected by the number of references of each of the seed publications (X), the pool publication (Y), and the shared references (Z). The number of shared references (Z) of each pool publication and the seed publication are identified manually. The labeling variable is a dichotomous variable: whether each pair of publications is similar (1 for similar pairs and 0 for dissimilar pairs). It is well known that the number of references of individual publications varies significantly, from a few to tens or hundreds. To make it so that the selected publication is able to represent publications with different numbers of references, and thus to ensure the estimation accuracy and application scope of the model, we carried out five steps. Step 1 was to arrange the number of references of each seed and each pool publication in descending order and exclude those with fewer than two references. Step 2 was to classify publications on the basis of their number of references into subset 1 (2-100 references), subset 2 (101-200 references), etc. The determination of the interval of the number of references is empirical. Step 3 was to randomly select publications from each subset, and in the end, a total of 500 sample publications from the seed and pool publication set, respectively, were obtained. Step 4 was to pair-match each of the sample publications in the seed set with each of those in the pool set, whereby a set of 250,000 pairs was obtained ($500 \times 500 = 250,000$). Step 5 was to arrange the number of shared references of each pair in descending order after excluding those with fewer than two shared references, to classify publication pairs on the basis of their number of shared references into subset 1 (2-20 shared references), subset 2 (21-40 shared references), etc., and to select randomly a total of 500 publication pairs from the subsets.

Of the 500 publication pairs, the mean values of the number of references of the seed and the pool publications are 89.6 and 97.0, respectively, while the corresponding modes are 62 and 61, respectively (Table 4). In combination with the results of the analysis of extreme value and standard deviation, one can see that the mean values are significantly affected by extreme values, and the data sampling is relatively dispersed, which is conducive to improving the accuracy of the

subsequent model estimation. More specifically, the samples were divided into two groups according to similar pairs (22.0%) and dissimilar pairs (78.0%) of publication pairs. The mean values of X and Y of the similar publication pairs are greater than those of the dissimilar pairs, which implies that a publication pair with a larger number of references may have more shared references (similar pairs: $Z_{\text{mean}} = 20.5$) and thus be more likely to be judged as a similar publication pair.

Table 4. Descriptive statistics on the number of references of 500 publication pairs

Statistics	Total samples			Dissimilar pairs (78.0%)			Similar pairs (22.0%)		
	X	Y	Z	X	Y	Z	X	Y	Z
Mean	89.6	97.0	9.9	79.5	89.8	6.9	125.7	122.4	20.5
Median	53.0	56.5	3.0	53.0	55.0	3.0	67.0	58.0	12.0
Mode	62	61	2	62	61	2	11	34	5
Standard deviation	107.3	114.9	14.5	97.0	108.1	11.0	131.8	133.7	19.7
Minimum	2	4	2	2	5	2	3	4	2
Maximum	503	660	81	503	660	72	502	482	81

Significant correlations between different variables do exist but with low values (less than 0.6, see Table 5). The correlation between X and Y (0.55) is higher than that between Z and X, Y (0.514 and 0.453, respectively). The correlations of the dependent variable with X and Y (0.093 and 0.041, respectively) were much lower than that with Z (0.488).

Table 5. Spearman's correlations of variables

Variables	X	Y	Z	Similarity
X	1.000			
Y	0.550** (0.000)	1.000		
Z	0.514** (0.000)	0.453** (0.000)	1.000	
Similarity	0.093* (0.037)	0.041 (0.360)	0.488** (0.000)	1.000

Note: *5% and **1% significance levels. The numbers shown in parentheses correspond to p -values.

7. Model 1: support vector machine

The support vector machine (SVM) is a supervised machine learning model for data classification that works by identifying the optimal hyperplane on the basis of the principle of maximum interval in the feature space (Cortes; Vapnik, 1995). Compared with methods such as setting percentage or quantity thresholds, this model can capture high-dimensional data features effectively. Since its inception, the SVM has been widely used in areas such as facial recognition, text classification, data classification, etc. (Qin; He, 2005; Sun *et al.*, 2002; Chen *et al.*, 2001; Campbell *et al.*, 2006; Srivastava; Bhambhu, 2010). This study proposes to approach the data binary classification issue by using the SVM model. We divided the sample data into two parts, consisting of the training data (400 publication pairs) for estimating the model parameters, and the testing data (100 publication pairs) for evaluating the model accuracy. The training and test data were randomly selected in a ratio of 4:1 for model fitting for 10 times. The given training samples ($D = \{(x_1, y_1), (x_2, y_2), \dots, (x_s, y_s)\}$, where s is the number of samples) are put into either negative or positive categories as labeled [$y \in (-1, 1)$]. The model can identify the optimal hyperplane, which requires maximizing the sum of sample-plane distances by dynamically adjusting the hyperplane parameters after placing all the training samples onto both sides of the hyperplane. The expression for the hyperplane is

$$H = wx + b = 0, x = (x_1, x_2, \dots, x_s) \quad (1)$$

The normal vector of the hyperplane is denoted as $w = (w_1, w_2, \dots, w_s)$, and the displacement as b . Finding the optimal hyperplane can be transformed into an optimization problem:

$$\begin{aligned} \min & \frac{|w|^2}{2} \\ \text{s.t. } & y_i [(w^T x_i) + b] \geq 1 \quad i = 1, 2, \dots, s \end{aligned} \quad (2)$$

As completely linearly separable cases are comparatively rare in reality, we can lower the standard for the support vector machine from completely linearly separable to tolerate slight classification errors by adding a slack variable $\xi_i (\xi_i \geq 0)$ and a penalty term $C (C > 0)$ to the original linearly separable support vector machine. The optimization problem can be solved by:

$$\begin{aligned} \min & \frac{|w|^2}{2} + C \sum_{i=1}^l \xi_i \\ \text{s.t. } & y_i [(w^T x_i) + b] \geq 1 - \xi_i, i = 1, 2, \dots, s \\ & \xi_i \geq 0, i = 1, 2, \dots, s \end{aligned} \quad (3)$$

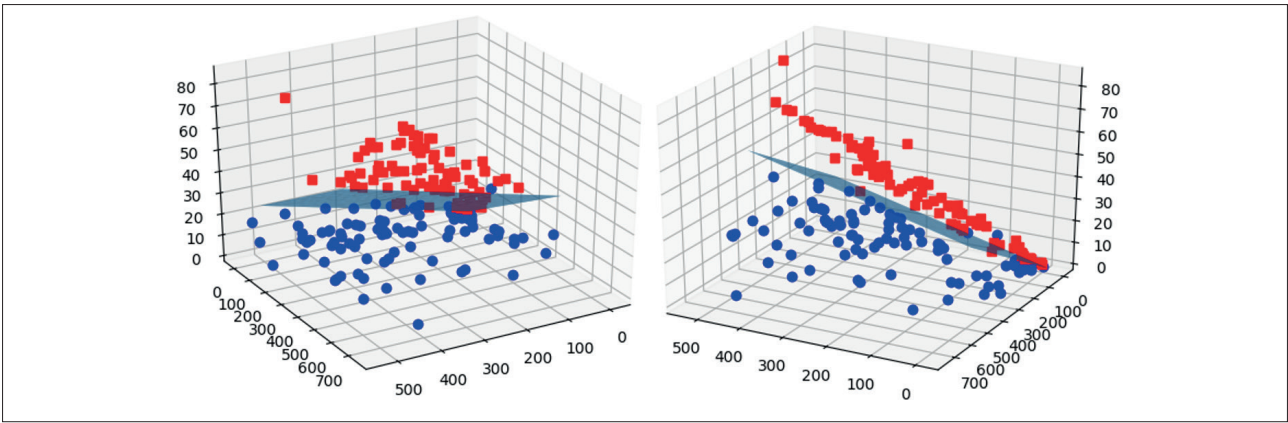


Figure 4. Support vector machine model

The model has a higher level of fault tolerance when $C = 1$. To solve Eq. (3), the Lagrange function can be constructed by adopting the Lagrange multiplier method, then the optimal solution to Eq. (3) can be obtained by solving its dual problem:

$$\begin{aligned} \min \frac{1}{2} \sum_{i=1}^s \sum_{j=1}^s \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^s \alpha_i \\ \text{s.t. } \sum_{i=1}^n \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C, i = 1, 2, \dots, s \end{aligned} \quad (4)$$

The sequential minimal optimization (SMO) algorithm can then be used to solve for α . Subsequently, w is solved via the equation $w = \sum_{i=1}^n \alpha_i y_i x_i$. The final function is:

$$f(x) = w^T x + b = \sum_{i=1}^n \alpha_i y_i x_i^T x_j + b \quad (5)$$

The hyperplane fitting result is shown in Figure 4. (The data presented in Figure 4 are the training samples.)

8. Model 2: fully connected neural network

A fully connected neural network (FCNN) is often used in data classification, information recognition, and prediction because of its strong self-learning and fitting abilities. It can fully approximate any complex nonlinear relationship (Liu et al., 2018; Raiyani et al., 2018). The FCNN constructed in the current study contains an input layer, an output layer, and two hidden layers. The training and testing datasets are the same as for the SVM. The input of each layer can be regarded as the output of the previous layer. The process of network generation consists of two steps in opposite directions, that is, forward propagation to calculate the model loss and backward propagation to update the model parameters. Finally, the model parameter value under the condition of minimum loss is obtained (Figure 5).

The forward propagation process of the network is to take the vector $x = [x_1, x_2, x_3]^T$ as the input data and to combine it linearly with the weights w_1, w_2, w_3 , and the bias term b . The nonlinear transformation is achieved by a linear rectification linear unit (ReLU) function, and the output vector z^1 is obtained and then used as the input data for the first hidden layer:

$$z^1 = f_1(W^1 x + B^1) \quad (6)$$

where $W^1 \in R^{3 \times n}$, $B^1 \in R^n$, and n is the number of neurons (32 neurons in this study) in this layer. The output vector of the first hidden layer can be regarded as the input vector of the second hidden layer. The formula for the second hidden layer is as follows:

$$z^2 = f_2(W^2 z^1 + B^2) \quad (7)$$

where $W^2 \in R^{n \times m}$ is the output weight matrix, $B^2 \in R^m$ is the bias vector, and m is the number of neurons (16 neurons in this study) in this layer. To prevent the model from overfitting and to improve the generalization ability of the network, a dropout layer is added to randomly stop updating the weights of some neurons (20% of neurons). The output vector of the second hidden layer is then used as the input vector of the output layer:

$$\hat{y} = f_3(W^3 z^2 + B^3) \quad (8)$$

where $W^3 \in R^{m \times 2}$, $B^3 \in R^2$. The network can be regarded as a nonlinear composite function:

$$\hat{y} = f(x, W, B) = f_3(W^3 f_2(W^2 f_1(W^1 x + B^1) + B^2) + B^3) \quad (9)$$

Through a softmax function, the output vector is transformed into two probability values (between 0 and 1), and the larger one can be converted into the prediction value 0 or 1 for judging the similarity between the seed publications and

pool publications. The difference between the values of prediction (\hat{y}) and ground truth (y) is calculated by a loss function and can be used as the criterion for model convergence.

The process of backpropagation is applied to optimize the model through the gradient descent algorithm. Common gradient descent algorithms include full gradient descent (FGD) and batch gradient descent (BGD) with relatively stable parameter updating direction but slow convergence speed (Huo; Huang, 2017). Stochastic gradient descent (SDG) improves the convergence speed of the first two methods by selecting training samples randomly in each iteration, but has the disadvantage of unstable parameter updating (Bottou, 2012). The adaptive gradient algorithm (AdaGrad) can obtain efficient and relatively correct results by constantly adjusting the learning rate in the iterative process (η), although calculation errors caused by the sharp decrease of the gradient are hard to avoid (Liu et al., 2020). In the current study, an optimization algorithm using root-mean-square propagation (RMSprop) of AdaGrad is used to construct the prediction model with a learning rate of 0.001 (Duchi et al., 2011; Dauphin et al., 2015). The formula is as follows:

$$\begin{aligned}
 g &\leftarrow \partial f(x^i; \theta), y^i \\
 r &\leftarrow \rho r + (1 - \rho)g \odot g \\
 \Delta \theta &= -\frac{\eta}{\sqrt{\delta + r}} \odot g \\
 \theta &\leftarrow \theta + \Delta \theta
 \end{aligned}
 \tag{10}$$

where δ , ϑ , ρ , and r are the constant value, model parameter, the exponential decay rate, and the gradient accumulation, respectively. The final training model is obtained with 97.5% prediction accuracy on the basis of the test data.

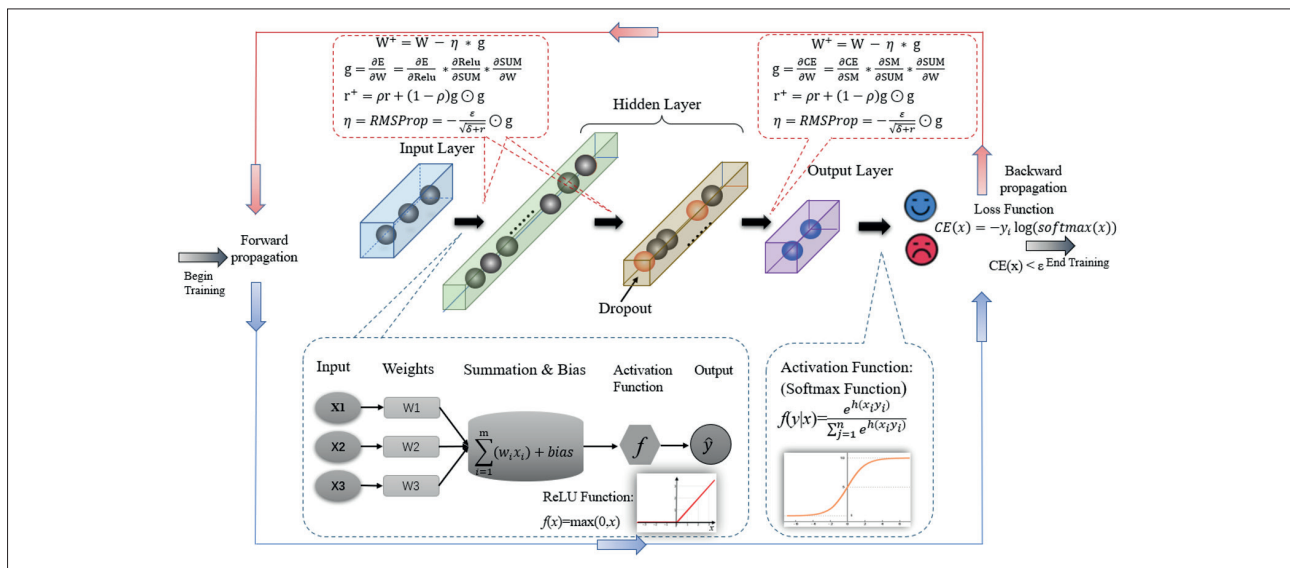


Figure 5. Fully connected neural network

9. Constructing two sets of expanded publications

Before running the above models, the follow steps are required to identify the shared references. The first step is to extract the shared references by comparing the DOIs of each reference pair between the seed publications and the pool publications. For references that cannot be matched by DOIs, other information in the CR field (i.e., author, year, journal, volume, and issue) can be used for fuzzy matching by adopting methods such as ignorance of punctuation differences, N-gram, and cosine (Abdulhayoglu et al., 2016; Glanzel; Czerwon, 1996; Sen; Gan, 1983). The current study adopts the method of ignorance of punctuation differences. By running the above models (SVM and FCNN), two sets of expanded publications are obtained (Figure 6). The training of the SVM or FCNN was performed on a NVIDIA GeForce GTX 1070 with an AMD Ryzen 1700. Sklearn 1.0.2 and Tensorflow 2.5.0 were used for the SVM and FCNN, respectively. For a computer with a memory of less than 500M, it takes less than 60 seconds to complete the training for each of the models.

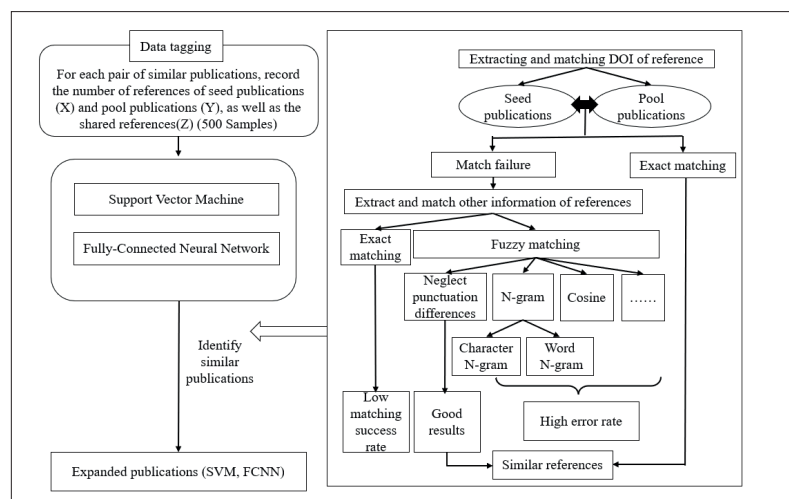


Figure 6. The process of constructing two sets of expanded publications

10. Comparison of the expanded collections obtained by the different methods

The three sets of expanded publications obtained by using the methods of Rakas and Hain (2019), SVM, and FCNN are shown in Figure 7. Overlaps exist between any two datasets, but with no full containment relationship. With 88,221 documents, the coverage of the method of Rakas and Hain is the widest, followed successively by that of SVM (22,138) and FCNN (15,925).

The accuracy and comprehensiveness of the three sets are presented in Table 6. The evaluation report of the models (e.g., accuracy, precision, recall, *f1*-score) and topic similarity (i.e., Jaccard, Word2vec, and TF-IDF) represent the fitting effect of each model and the semantic similarity between each seed document and its expanded documents, calculated according to the information in the title, abstract, keywords, and keywords plus. The results from the method of Rakas and Hain are unsatisfactory because of the large number of duplicate publications and the lowest average topic similarity (TF-IDF = 10.4%). The SVM exhibits relatively low accuracy (87.5%), recall (79.2%), and *f1*-score (82.7%). The FCNN performs best, with the highest model accuracy (97.5%), precision (96.2%), recall (98.2%), *f1*-score (97.5%), and average topic similarity (TF-IDF = 25.1%). Thus, the results of the FCNN model are incorporated into the expanded set.

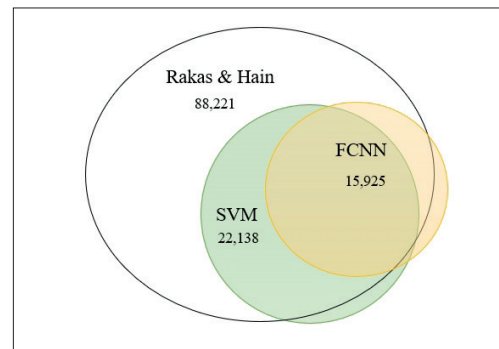


Figure 7. Coverage of the three sets of expanded publications

Table 6. Results of the three publication expansion models

Method	Expanded publications	Expanded publications (De-duplicate)	Model evaluation				Topic similarity		
			Accuracy	Precision	Recall	<i>f1</i> -score	Jaccard	Word2vec	TF-IDF
Rakas & Hain	1,668,408	88,221	-	-	-	-	4.0%	9.3%	10.4%
Support Vector Machine (Linear kernel function)	88,885	22,138	95.00%	94.05%	94.05%	94.05%	6.6%	18.4%	19.7%
Fully-Connected Neural Network	47,189	15,925	97.50%	96.15%	98.21%	97.53%	8.7%	21.9%	25.1%

11. Publication expansion based on topic similarity

Selecting expanded publications simply on the basis of shared references may neglect publication pairs that do not have enough shared references but do study similar topics. Therefore, we apply both shared references and topic similarity to screen out publications in the pool set. In other words, publications in the expanded set are screened out firstly by shared references and then by topic similarity. The topic similarity method is a two-step process. Step 1 is to calculate the topic similarity between the expanded publications obtained through shared references and their corresponding seed publications. The value of 25% (in line with Table 6) will be used to screen out publication pairs in step 3. Step 2 is to pair the rest of the publications (excluding those already screened out through shared references) in the pool set with those in the seed set, and calculate the TF-IDF value of each pair. Step 3 is to screen out those with topic similarity greater than 25% and add them to the expanded set.

By integrating the expanded publications on the basis of shared references and topic similarity, an expanded set (Table 7) containing 50,681 publications is obtained after removing duplicates, with 45 subject categories, among which environmental studies (12.6%), environmental sciences (10.4%), and economics (9.8%) account for a large proportion. The rest of the categories include health care sciences and services, education and educational research, political science,

Table 7. Distribution of the top 20 disciplines in the expanded set

Discipline	N	Ratio
Environmental studies	13,562	12.6%
Environmental sciences	11,218	10.4%
Economics	10,581	9.8%
Public administration	9,455	8.8%
Health policy and services	5,153	4.8%
Management	4,911	4.6%
Healthcare sciences and services	4,798	4.5%
Education and educational research	3,788	3.5%
Energy and fuels	3,767	3.5%
Political science	3,417	3.2%
Public, environmental, and occupational health	3,221	3.0%
Information science and library science	3,089	2.9%
Business	2,599	2.4%
Geography	2,150	2.0%
Ecology	1,775	1.6%
Urban studies	1,682	1.6%
Social issues	1,614	1.5%
Industrial relations and labor	1,475	1.4%
Computer science, interdisciplinary applications	1,402	1.3%
Regional and urban planning	1,287	1.2%

public, environmental and occupational health, information science and library science, etc. (Table 7). Compared with the pool set, the importance of some disciplines in this set has increased, for example, public administration (8.8%), health policy and services (4.8%), and management (4.6%).

12. Step 4. Construct a final corpus of publications

Finally, the seed publications and expanded publications are merged together by excluding duplicates, reviews, and uncited publications (Figure 8). The final corpus contains a total of 55,345 publications, including 45 WoS subject categories and 116 journals and research relating to scientific research management and innovation research, public health and health, safety and risk management, organizational management and human resource management, land resource management, water resource management, climate change, energy and environmental governance, governance research, education, social issues research, and other content connected to public affairs.

In short, the advantages of the multistep method are obvious. The seed set contains the core content of PA, while the expanded set considers its dynamic expansion. Thus, as a result of inclusiveness, it will not provide a specific boundary for PA, but rather a blurry one that tends to include publications from adjacent fields that are also substantially interrelated to PA.

13. Discussion and conclusions

To solve the problem of how to retrieve publications in interdisciplinary research, this study proposes a data-driven and multi-round iterative retrieval strategy that may ameliorate the deviation of retrieval results caused by single-way methods such as subject retrieval or journal retrieval. Selecting a reasonable method for expanding publication coverage is critical for the retrieval strategy, directly affecting the accuracy of the final retrieved results. Compared with the methods of setting an absolute number or relative threshold (e.g., Rakas; Hain, 2019) for expanding coverage, the FCNN model performs better because it addresses the difficulty of comparing the similarity of publications caused by large differences in the number of references. Furthermore, it improves the accuracy of determining similar publications. The proposed search strategy can be applied for the retrieval of any interdisciplinary publications. Adjustment, however, has to be done according to the characteristics of the target disciplines.

The third step of the search strategy—establishing an expanded set of publications—is somewhat complicated and requires knowledge of computer science and bibliometrics, which may, to some extent, limit the application of this methodology. To solve this problem, we plan to develop user-friendly guidelines and to make them freely accessible, so that researchers in fields with little knowledge of computer science and bibliometrics may apply the guidelines to collect an expanded set of publications in the interdisciplinary area of their choosing.

14. References

- Abdulhayoglu, Mehmet-Ali; Thijs, Bart; Jeuris, Wouter** (2016). "Using character n-grams to match a list of publications to references in bibliographic databases". *Scientometrics*, v. 109, n. 3, pp. 1525-1546.
<https://doi.org/10.1007/s11192-016-2066-3>
- Alencar, Maria-Simone-de-M.; Porter, Alan L.; Antunes, Adelaide-Maria-de-Souza** (2007). "Nanopatenting patterns in relation to product life cycle". *Technological forecasting and social change*, v. 74, n. 9, pp. 1661-1680.
<https://doi.org/10.1016/j.techfore.2007.04.002>
- Archambault, É.; Beauchesne, Olivier H.; Caruso, Julie** (2011). "Towards a multilingual, comprehensive and open scientific journal ontology". In: *Proceedings of the 13th International conference of the International Society for Scientometrics and Informetrics*, pp. 66-77. South Africa, Durban.
- Bassecouard, Elise; Lelu, Alain; Zitt, Michel** (2007). "Mapping nanosciences by citation flows: A preliminary analysis". *Scientometrics*, v. 70, n. 3, pp. 859-880.
<https://doi.org/10.1007/s11192-007-0315-1>
- Benz, Arthur** (2005). "Public administrative science in Germany: Problems and prospects of a composite discipline". *Public administration*, v. 83, n. 3, pp. 659-668.
<https://doi.org/10.1111/j.0033-3298.2005.00468.x>
- Bottou, Léon** (2012). *Stochastic gradient descent tricks*. Berlin, Heidelberg: Springer.
https://link.springer.com/chapter/10.1007/978-3-642-35289-8_25
- Boyack, Kevin W.; Klavans, Richard** (2010). "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?". *Journal of the American Society for Information Science and Technology*, v. 61, n. 12, pp. 2389-2404.
<https://doi.org/10.1002/asi.21419>

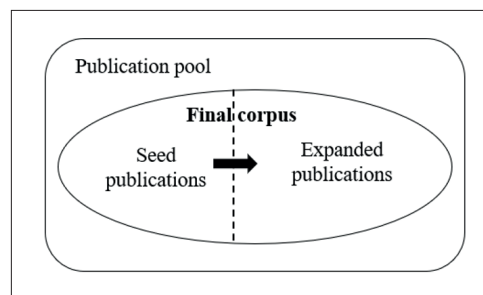


Figure 8. Relationship between different publication sets

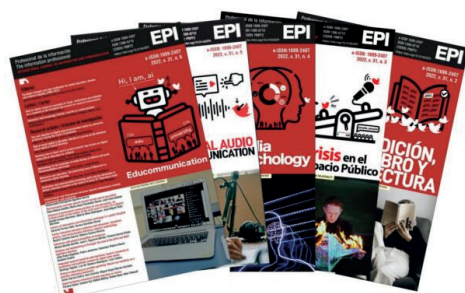
- Campbell, William M.; Sturim, Douglas E.; Reynolds, Douglas A.; Solomonoff, Alex** (2006). "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation". In: *2006 IEEE International conference on acoustics speech and signal processing proceedings*, v. 1, pp. I-I.
- Chen, Yunqiang; Zhou, Xiang-Sean; Huang, Thomas S.** (2001). "One-class SVM for learning in image retrieval". In: *Proceedings 2001 international conference on image processing*, v. 1, pp. 34-37.
<https://doi.org/10.1109/ICIP.2001.958946>
- Cortes, Corinna; Vapnik, Vladimir** (1995). "Support-vector networks". *Machine learning*, v. 20, n. 3, pp. 273-297.
<https://doi.org/10.1007/BF00994018>
- Dauphin, Yan N.; De-Vries, Harm; Bengio, Yoshua** (2015). "Equilibrated adaptive learning rates for non-convex optimization". *Advances in neural information processing systems*.
- Denhardt, Robert B.** (2004). *Theories of public organization*. Belmont: Wadsworth.
- Duchi, John C.; Hazan, Elad; Singer, Yoram** (2011). "Adaptive subgradient methods for online learning and stochastic optimization". *Journal of machine learning research*, v. 12, n. 61, pp. 2121-2159.
<https://jmlr.org/papers/v12/duchi11a.html>
- Fagerberg, Jan; Fosaas, Morten; Sappasert, Koson** (2012). "Innovation: Exploring the knowledge base". *Research policy*, v. 41, n. 7, pp. 1132-1153.
<https://doi.org/10.1016/j.respol.2012.03.008>
- Fleisher, Craig S.; Blair, Natasha M.** (1999). "Tracing the parallel evolution of public affairs and public relations: An examination of practice, scholarship and teaching". *Journal of communication management*, v. 3, n. 3, pp. 276-292.
<https://doi.org/10.1108/eb026052>
- Glänzel, Wolfgang** (2015). "Bibliometrics-aided retrieval: where information retrieval meets scientometrics". *Scientometrics*, v. 102, n. 3, pp. 2215-2222.
<https://doi.org/10.1007/s11192-014-1480-7>
- Glänzel, Wolfgang; Czerwon, Hans-Jürgen** (1996). "A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level". *Scientometrics*, v. 37, n. 2, pp. 195-221.
<https://doi.org/10.1007/bf02093621>
- Glänzel, Wolfgang; Debackere, Koenraad** (2021). "Various aspects of interdisciplinarity in research and how to quantify and measure those". *Scientometrics*, v. 127, n. 9, pp. 5551-5569.
<https://doi.org/10.1007/s11192-021-04133-4>
- Glänzel, Wolfgang; Schubert, András** (2003). "A new classification scheme of science fields and subfields designed for scientometric evaluation purposes". *Scientometrics*, v. 56, n. 3, pp. 357-367.
<https://doi.org/10.1023/A:1022378804087>
- Gómez-Núñez, Antonio J.; Batagelj, Vladimir; Vargas-Quesada, Benjamín; De-Moya-Anegón, Félix; Chinchilla-Rodríguez, Zaida** (2014). "Optimizing SCImago Journal & Country Rank classification by community detection". *Journal of informetrics*, v. 8, n. 2, pp. 369-383.
<https://doi.org/10.1016/j.joi.2014.01.011>
- Griffith, Berver C.; Small, Henry G.; Stonehill, Judith A.; Dey, Sandra** (1974). "The structure of scientific literatures II: Toward a macro- and microstructure for science". *Science studies*, v. 4, n. 4, pp. 339-365.
<https://doi.org/10.1177/030631277400400402>
- Harris, Phil; Fleisher, Craig S.** (2005). *The handbook of public affairs*. Sage.
<https://doi.org/10.4135/9781848608108>
- Harris, Phil; Moss, Danny** (2001a). "In search of public affairs: a function in search of an identity". *Journal of public affairs*, v. 1, n. 2, pp. 102-110.
<https://doi.org/10.1002/pa.55>
- Harris, Phil; Moss, Danny** (2001b). "Understanding public affairs". *Journal of public affairs*, v. 1, n. 1, pp. 6-8.
<https://doi.org/10.1002/pa.45>
- Hicks, Diana; Wouters, Paul; Waltman, Ludo; De-Rijcke, Sarah; Ràfols, Ismael** (2015). "Bibliometrics: The Leiden Manifesto for research metrics". *Nature*, v. 520, n. 7548, pp. 429-431.
<https://doi.org/10.1038/520429a>
- Huang, Can; Notten, Ad; Rasters, Nico** (2010). "Nanoscience and technology publications and patents: a review of social science studies and search strategies". *The journal of technology transfer*, v. 36, n. 2, pp. 145-172.
<https://doi.org/10.1007/s10961-009-9149-8>

- Huang, Mu-Hsuan; Chang, Chia-Pin** (2013). "Detecting research fronts in OLED field using bibliographic coupling with sliding window". *Scientometrics*, v. 98, n. 3, pp. 1721-1744.
<https://doi.org/10.1007/s11192-013-1126-1>
- Huang, Ying; Schuehle, Jannik; Porter, Alan L.; Youtie, Jan** (2015). "A systematic method to create search strategies for emerging technologies based on the Web of Science: illustrated for 'Big Data'". *Scientometrics*, v. 105, n. 3, pp. 2005-2022.
<https://doi.org/10.1007/s11192-015-1638-y>
- Huo, Zhouyuan; Huang, Heng** (2017). "Asynchronous mini-batch gradient descent with variance reduction for non-convex optimization". In: *Proceedings of the AAAI Conference on artificial intelligence*, v. 31, n. 1.
<https://doi.org/10.1609/aaai.v31i1.10940>
- Kettl, Donald F.; Milward, H. Brinton** (1996). *The state of public management*. Baltimore: Johns Hopkins University Press.
- Klein, Julie-Thompson; Newell, William H.** (1996). "Advancing interdisciplinary studies". In: J. G. Gaff; J. L. Ratcliff; Associates (eds.). *Handbook of the undergraduate curriculum. A comprehensive guide to purposes, structures, practices, and change*, pp. 393-415. San Francisco, CA: Jossey-Bass.
- Kostoff, Ronald N.; Murday, James S.; Lau, Clifford G. Y.; Tolles, William M.** (2005). "The seminal literature of nanotechnology research". *Journal of nanoparticle research*, v. 8, n. 2, pp. 193-213.
<https://doi.org/10.1007/s11051-005-9034-9>
- Ledford, Heidi** (2015). "How to solve the world's biggest problems". *Nature*, v. 525, n. 7569, pp. 308-311.
<https://doi.org/10.1038/525308a>
- Lerbinger, Otto** (2006). *Corporate public affairs: Interacting with interest groups, media, and government*. Routledge. ISBN: 978 0 805856439
- Leydesdorff, Loet; Ràfols, Ismael** (2009). "A global map of science based on the ISI subject categories". *Journal of the American Society for Information Science and Technology*, v. 60, n. 2, pp. 348-362.
<https://doi.org/10.1002/asi.20967>
- Leydesdorff, Loet; Wagner, Caroline S.; Bornmann, Lutz** (2019). "Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient". *Journal of informetrics*, v. 13, n. 1, pp. 255-269.
<https://doi.org/10.1016/j.joi.2018.12.006>
- Liu, Kui; Kang, Guixia; Zhang, Ningbo; Hou, Beibei** (2018). "Breast cancer classification based on fully-connected layer first convolutional neural networks". *IEEE Access*, v. 6, pp. 23722-23732.
<https://doi.org/10.1109/ACCESS.2018.2817593>
- Liu, Mingrui; Mroueh, Youssef; Ross, Jerret; Zhang, Wei; Cui, Xiaodong; Das, Payel; Yang, Tianbao** (2020). "Towards better understanding of adaptive gradient algorithms in generative adversarial nets". *arXiv*, arXiv:1912.11940.
- Liu, Na; Shapira, Philip; Yue, Xiaoxu** (2021). "Tracking developments in artificial intelligence research: constructing and applying a new search strategy". *Scientometrics*, v. 126, n. 4, pp. 3153-3192.
<https://doi.org/10.1007/s11192-021-03868-4>
- Lovelock, James E.; Margulis, Lynn** (1974). "Atmospheric homeostasis by and for the biosphere: the gaia hypothesis". *Tellus*, v. 26, n. 1-2, pp. 2-10.
<https://doi.org/10.1111/j.2153-3490.1974.tb01946.x>
- Mogoutov, Andrei; Kahane, Bernard** (2007). "Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking". *Research policy*, v. 36, n. 6, pp. 893-903.
<https://doi.org/10.1016/j.respol.2007.02.005>
- Muñoz-Écija, Teresa; Vargas-Quesada, Benjamín; Chinchilla-Rodríguez, Zaida** (2019). "Coping with methods for delineating emerging fields: Nanoscience and nanotechnology as a case study". *Journal of informetrics*, v. 13, n. 4, pp. 100976-100976.
<https://doi.org/10.1016/j.joi.2019.100976>
- Organisation for Economic Cooperation and Development* (1972). *Interdisciplinarity: Problems of teaching and research in universities*. Centre for Educational Research and Innovation. OECD Publications Center. Paris, France.
- Ostrom, Vincent** (1974). *The intellectual crisis in American public administration*. University, Ala., University of Alabama Press.
- Qin, Jun; He, Zhong-Shi** (2005). "A SVM face recognition method based on Gabor-featured key points". In: *2005 International conference on machine learning and cybernetics*, v. 8, pp. 5144-5149.

- Raadschelders, Jos C. N.** (2011). *Public administration: The interdisciplinary study of government*. Oxford/New York: Oxford University Press.
- Raiyani, Kashyap; Gonçalves, Teresa; Quaresma, Paulo; Nogueira, Vitor-Beires** (2018). "Fully connected neural network with advance preprocessor to identify aggression over Facebook and Twitter". In: *Proceedings of the first workshop on trolling, aggression and cyberbullying*, pp. 28-41.
<https://www.semanticscholar.org/paper/Fully-Connected-Neural-Network-with-Advance-to-over-Raiyani-Gon%C3%A7alves/ad0679c71836e32f3edf6f0234b9a2f486b132ed>
- Rakas, Marija; Hain, Daniel S.** (2019). "The state of innovation system research: What happens beneath the surface?". *Research policy*, v. 48, n. 9, 103787.
<https://doi.org/10.1016/j.respol.2019.04.011>
- Repko, A.; Nacakas, Francine; Fiscella, Joan B.** (2007). "Integrating Interdisciplinary: How the theories of common ground and cognitive interdisciplinarity are informing the debate on interdisciplinary integration". *Issues in interdisciplinary studies*, n. 25, 31 pp.
<https://www.semanticscholar.org/paper/Integrating-Interdisciplinary-Ity%3A-How-the-Theories-Repko-Navakas/b8c2f4e6d63c030068bbb93b65df408748a6b54d>
- Rodgers, Robert; Rodgers, Nanette** (2000). "Defining the boundaries of public administration: Undisciplined mongrels versus disciplined purists". *Public administration review*, v. 60, n. 5, pp. 435-445.
<https://doi.org/10.1111/0033-3352.00106>
- Rotolo, Daniele; Hicks, Diana; Martin, Ben R.** (2015). "What is an emerging technology?". *Research policy*, v. 44, n. 10, pp. 1827-1843.
<https://doi.org/10.1016/j.respol.2015.06.006>
- Ruiz-Navas, Santiago; Miyazaki, Kumiko** (2018). "A complement to lexical query's search-term selection for emerging technologies: the case of 'big data'". *Scientometrics*, v. 117, n. 1, pp. 141-162.
<https://doi.org/10.1007/s11192-018-2857-9>
- Savage, Peter** (1974). "Dismantling the administrative state: Paradigm reformulation in public administration". *Political studies*, v. 22, n. 2, pp. 147-157.
<https://doi.org/10.1111/j.1467-9248.1974.tb00014.x>
- Seibel, Wolfgang** (1996). "Administrative science as reform: German public administration". *Public administration review*, v. 56, n. 1, p. 74.
<https://doi.org/10.2307/3110057>
- Sen, Subir K.; Gan, Shymal K.** (1983). "A mathematical extension of the idea of bibliographic coupling and its applications". *Annals of library science and documentation*, v. 30, n. 2, pp. 78-82.
[https://nopr.niscpr.res.in/bitstream/123456789/28008/1/ALIS%2030\(2\)%2078-82.pdf](https://nopr.niscpr.res.in/bitstream/123456789/28008/1/ALIS%2030(2)%2078-82.pdf)
- Shapira, Philip; Kwon, Seokbeom; Youtie, Jan** (2017). "Tracking the emergence of synthetic biology". *Scientometrics*, v. 112, n. 3, pp. 1439-1469.
<https://doi.org/10.1007/s11192-017-2452-5>
- Small, Henry; Griffith, Belver C.** (1974). "The structure of scientific literatures I: Identifying and graphing specialties". *Science studies*, v. 4, n. 1, pp. 17-40.
<https://doi.org/10.1177/030631277400400102>
- Small, Henry; Sweeney, Ed; Greenlee, Edward** (1985). "Clustering the science citation index using co-citations. II. Mapping science". *Scientometrics*, v. 8, n. 5-6, pp. 321-340.
<https://doi.org/10.1007/bf02018057>
- Soós, Sándor; Kiss, Anna** (2020). "Informetrics and the study of science-society communications: a bibliometric scoping review". *Scientometrics*, v. 124, n. 2, pp. 825-842.
<https://doi.org/10.1007/s11192-020-03444-2>
- Srivastava, Durgesh, K.; Bhambhu, Lekha** (2010). "Data classification using support vector machine". *Journal of theoretical and applied information technology*, v. 12, n. 1.
- Stallings, Robert A.** (1986). "Doctoral programs in public administration: An outsider's perspective". *Public administration review*, v. 46, n. 3, p. 235.
<https://doi.org/10.2307/3110438>
- Steckmest, Francis W.; Business Roundtable** (1982). *Corporate performance: The key to public trust*. New York: McGraw-Hill.

- Sun, Aixin; Lim, Ee-Peng; Ng, Wee-Keong** (2002). "Web classification using support vector machine". In: *Proceedings of the fourth international workshop on Web information and data management - WIDM'02*.
<https://doi.org/10.1145/584931.584952>
- Suominen, Arho; Li, Yin; Youtie, Jan; Shapira, Philip** (2016). "A bibliometric analysis of the development of next generation active nanotechnologies". *Journal of nanoparticle research*, v. 18, n. 9.
<https://doi.org/10.1007/s11051-016-3578-8>
- Wagner, Caroline S.; Roessner, J. David; Bobb, Kamau; Klein, Julie-Thompson; Boyack, Kevin W.; Keyton, Joann; Ràfols, Ismael; Börner, Katy** (2011). "Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature". *Journal of informetrics*, v. 5, n. 1, pp. 14-26.
<https://doi.org/10.1016/j.joi.2010.06.004>
- Waltman, Ludo; Van-Eck, Nees-Jan** (2012). "A new methodology for constructing a publication-level classification system of science". *Journal of the American Society for Information Science and Technology*, v. 63, n. 12, pp. 2378-2392.
<https://doi.org/10.1002/asi.22748>
- Wang, Jian; Thijs, Bart; Glänzel, Wolfgang** (2015). "Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity". *PLoS one*, v. 10, n. 5, e0127298.
<https://doi.org/10.1371/journal.pone.0127298>
- Wang, Zhinan; Porter, Alan L.; Kwon, Seokbeom; Youtie, Jan; Shapira, Philip; Carley, Stephen F.; Liu, Xiaoyu** (2019). "Updating a search strategy to track emerging nanotechnologies". *Journal of nanoparticle research*, v. 21, n. 9.
<https://doi.org/10.1007/s11051-019-4627-x>
- White, Jay D.; Adams, Guy B.; Forrester, John P.** (1996). "Knowledge and theory development in public administration: The role of doctoral education and research". *Public administration review*, v. 56, n. 5, p. 441.
<https://doi.org/10.2307/977043>
- Xue Lan; Liang, Zheng; Yang, Liexun** (2017). *The development strategy of public management*. Science Press, Beijing.
- Yang, Lihua** (2019). "Public administration as a dynamic balance and integrative science across Politics, Management, and Law: Rosenbloom's framework and Chinese experiences". *The American review of public administration*, v. 49, n. 1, pp. 79-97.
<https://doi.org/10.1177/0275074018759337>
- Yu, Z.** (2019). *Research on the development and local construction of public administration discipline in China - a perspective of bibliometrics*. Unpublished doctoral dissertation, Zhejiang University, Hangzhou.
- Zhou, Xiao; Huang, Lu; Zhang, Yi; Yu, Miaomiao** (2019). "A hybrid approach to detecting technological recombination based on text mining and patent network analysis". *Scientometrics*, v. 121, n. 2, pp. 699-737.
<https://doi.org/10.1007/s11192-019-03218-5>
- Zitt, Michel; Bassecoulard, Elise** (2006). "Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences". *Information processing & management*, v. 42, n. 6, pp. 1513-1531.
<https://doi.org/10.1016/j.ipm.2006.03.016>

REVISTA EPI



<https://www.profesionaldelainformacion.com>

Profesional de la información es una revista científica sobre Comunicación, Documentación, Biblioteconomía, Informetría y Ciencias de la información. Se publica cada 2 meses (6 números al año).

e-ISSN: 1699-2407

<https://doi.org/10.3145/EPI>

SJR 2022 = 0,872 (Q1); JIF 2022 = 4,2 (Q1)