PLOS COMPLEX SYSTEMS

RESEARCH ARTICLE

Truth set size prediction by Newton's cooling law

Yuji Fujita[1,2☯]*, **Noritaka Usami**[2,3,4☯], **Toshiaki Fujii**[2,3☯], **Hiroaki Nagai**[2,4,5☯]

1 National Graduate Institute for Policy Studies (GRIPS), Tokyo, Japan, **2** Japan Cabinet Office, Tokyo, Japan (Affiliation of this study was conducted), **3** Nagoya University Graduate School of Engineering, Nagoya, Japan, **4** Nagoya University Academic Research & Industry-Academia-Government Collaboration (AR&IAGC), Nagoya, Japan, **5** Nagoya University IR Office, Nagoya, Japan

☯ These authors contributed equally to this work.
* yuji.fujita.3.14@gmail.com

Abstract

Precision and recall are useful indices to evaluate an operation, algorithm, system, and other subjects from two different facets. However, they are not readily available when the subject is still in progress because the truth set, which is required to calculate recall, is unknown. In this study, a method to predict the *size* of the truth set of an inquiry still in progress is presented, which consists of a classical 18th century mechanics found and formulated by Isaac Newton, today known as "Newton's cooling law", with some set-theoretical trick and executed by Markov Chain Monte Carlo. The developed method is applied to nation-wide scale collections of identifications of the authors of academic articles as the affiliation data of Japanese national research organizations, and obtain recall values, as a part of objective, evidence-based policy for science and technology of the Japanese government. The author identification result is naturally represented as a bipartite directed graph, from the set of authors to the set of affiliation data. We conduct a sort of network prediction, not on the bipartite graph itself but on its *vertices size* and obtain the true graph size by using a simple and straightforward probabilistic model, which is implemented by also a classical, yet recently developing probabilistic inference method.

Author summary

In this study, we propose a method to predict the value of recall of the unfinished work, for example, survey or inquiry still in progress. It may sound strange to discuss recall or precision on a subject still going on, yet, they are needed in the real world applications because we want to know how much more we can expect from the unfinished inquiry. The trick is to predict the *size* of the truth set, not the set itself. If we observe identical subject multiple times, the observations must be slightly different from each other, depending on how much left to be done. If the results are almost equal, the observations are good approximation of the truth set, and the recall should be high. Conversely, disagreement between the observations is a sign of low recall. The method consists of simple and classic early eighteenth century mechanics found and formulated by Sir Isaac Newton and

Citation: Fujita Y, Usami N, Fujii T, Nagai H (2024) Truth set size prediction by Newton's cooling law. PLOS Complex Syst 1(4): e0000020. https://doi.org/10.1371/journal.pcsy.0000020

Editor: Jaya Sreevalsan-Nair, International Institute of Information Technology Banggalore, INDIA

Received: March 22, 2024

Accepted: September 16, 2024

Published: December 5, 2024

Data Availability Statement: All data are in the manuscript and/or Supporting information files.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have no competing interests on this research.

known as "Newton's cooling law" today, some mathematical tricks for data preparation, and Markov Chain Monte Carlo. We apply the method to nation-wide scale collections of article authors identification as the affiliation data of Japanese national research organizations, and predict how many researchers are left to be identified as authors. As the identification relation from an author to the affiliation data of a researcher is naturally represented as a directed bipartite graph, a form of network-prediction on a directed bipartite graph is executed in this study.

Introduction

Precision and recall are two useful ways of quantitative evaluation of tests, operations, algorithms, systems, collection of information, and various other subjects. They are clearly defined and capable of capturing two different facets of the subject we wish to examine.

However, they are not always available in the real world applications. Especially the truth set, which is required to calculate the recall value, is basically unavailable in an excavation-like inquiry process. Nonetheless, precision and recall are seriously needed in such cases. We all wish to know how well the inquiry being conducted is going on, which is naturally evaluated by the proportion of good results in the work, which is precision, and the coverage of the result, which is recall.

In this study, we present a method to predict the unknown size of the truth set that is needed to obtain the recall value. The basic idea is as follows: Observe identical object multiple times. As observation is naturally imperfect, such observations will not be exactly equal. It is the difference between these observations that gives the key to estimating the size of the truth set. If the observations are significantly different from each other, there should be much left to be observed, and if the observations are approximately equal, there is not much left unobserved, in which case the recall value should be high. The method does not provide the truth set itself, but the *size* of it.

The method is realized as a simple and classic probabilistic model, which has its origin in the early eighteenth century mechanics found and formulated by Sir Isaac Newton as [1], today known as the Newton's Cooling Law. The estimation is executed by a probabilistic inference procedure known as Markov Chain Monte Carlo (MCMC hereafter), with some set-theoretical tricks for data preparation.

In this study, we apply our method to the author-identification result of academic papers in nation-wide scale. Actually the method was developed for the purpose of quantitative evaluation of our author-identification result.

Author identification is a process to identify the author of an academic paper as a researcher affiliated to some research organization. Such a process may look trivial when we find an acquainted researcher's name on a academic paper's cover page. However, the naive confidence abundant in our neighborhood is hardly kept if comprehensive, nation-wide collection of author-researcher references is to be obtained using research database products like Scopus or Dimensions and research organization's affiliation data.

Actually, identification is not trivially executed in general. In the famous short story of Sherlock Holmes "The Man with the Twisted Lip", one character St. Clair, and another character Hugh Boone are presented as two different men. The case is settled by the detective who demonstrates that these two characters actually refer to a single person. Identification is a typically inquiry-type process.

The authors, who are (or were) affiliated to the Council for Science, Technology, and Innovation, which is a department of the Cabinet Office, Government of Japan, are working on evidence-based policy-making of science policy, for which it is essential to measure the academic performance of national research organizations. For example [2] is a scale-invariant institutional citation index based on the principle of fractal developed by two of the authors of this study.

The author identification gives the foundation of various performance measurements. Because the career path of scientists is subject to change and diversification, measurement of academic performance is becoming progressively difficult (see [3]).

The identification process itself gained attention in these several decades as a topic of information and communications technology, where the question is referred to as a "reference problem" or "reference disambiguation"(see [4], [5] or more directly [6]). However, the novelty of this research does not exist there. This research is about the method of estimating truth set size in an inquiry-type process, where the truth set is not given.

The result of identification, or the reference from the author to the researcher, is naturally represented as a directed bipartite graph. In network science, there is a topic called *network prediction*, where the imperfection of observation is mitigated to reach the innate network structure (see [7] for description and review). In this research, we perform a form of network prediction, not on the bipartite graph itself but on its *size*.

Throughout this research, we will use the term "author" as the source vertex of the bipartite graph, found in the published academic research, and "researcher" as the destination vertex, a person to be identified as a researcher who belongs to a research organization.

Materials and methods

Basic preparation

Let X be a set of source vertices and Y be the set of destination vertices of a directed bipartite graph. In this research, the set X is the set of authors that can be found in a research database product, and Y is the set of researchers who belong to some national institute. When an author $x \in X$ is identified as a researcher $y \in Y$, an edge from x to y as $(x, y) \in E$ (let E be a set of edges) is formed. Thus, a directed bipartite graph G is specified by the set of edges E, and the source and destination vertices sets X, Y.

As there are several database products available in the market, they are indexed as X_i, $i = 1, 2, 3, \ldots$. By using different products, corresponding different collection of identifications, and associated bipartite graph G_i results.

As the purpose of this study is to evaluate quantitatively how good the obtained graph is, the precision and recall are formulated as follows: let G_∞ be the true bipartite graph, with corresponding vertices X_∞, Y_∞, and the edges E_∞. The term "true" means no researcher left unidentified, nor misidentified. Obviously such a result is unavailable because no database product or identification process is perfect.

Let Y_G be the referred researchers' set of particular graph G. Then, the precision is formulated as

$$\frac{|Y_\infty \cap Y_G|}{|Y_G|}, \tag{1}$$

which means the proportion of the referred researchers who actually wrote some academic

papers. Recall on the other hand is

$$\frac{|Y_\infty \cap Y_G|}{|Y_\infty|}, \tag{2}$$

which is the proportion of the referred researcher who became the author of some academic paper.

Data

Here we describe how the bipartite graphs, which are to be quantitatively evaluated by precision and recall, are obtained. The collection of identification E of the bipartite graph G is realized by two factors, the affiliation and the name (the first name and the family name) of the researcher. Suppose an author $x \in X$ has a certain name and affiliation. If we could find an entity $y \in Y$ who has the same name and affiliation, then we identify x as $y \in Y$ and count the pair as $(x, y) \in E$; set the edge from the author to the researcher. In this research, the set of the researchers Y is cleaned in a regular basis and managed to keep its soundness and integrity as a project of the Japanese government for the purpose of governmental research funding management. It is named "e-Rad" system (e-Rad hereafter); the data from e-Rad consists of the destination side of the graph. The set of researchers is given as the snapshot of e-Rad taken in November 2020, which consists of 1,079,535 records of 758,087 researchers from national universities, national research institutes, research departments of private companies, and other various organizations.

We have constructed four bipartite graphs from four different research database products, which are Dimensions by DigitalScience, Inc., Scopus, and its special extended database by JSTAGE from Elsevier associated with Japan Science and Technology Agency (JST), and Web of Science from Clarivate Plc. (in alphabetical order). The data used in this research are the snapshots of April 2022.

The difference of snapshot timestamps between the source and destination comes from the fact that it takes time for academic articles to be published, and even more time to be recorded in a database product, while e-Rad data is kept updated constantly.

In this research, there are four different bipartite graphs available, according to the research database product. The obtained bipartite graph's link size is from 285,388 to 393,279, the author (source vertices) set size is from 209,060 to 33,5147, and the researcher set size is 193,116 to 247,744. The source vertices set size differs more than the destination vertices set size due to the product vendor's own author identification policies.

Hence our method depends on the difference in the destination vertices size of the bipartite graph, we give further data description by showing the destination vertices size (the number of affiliated people identified as authors) of 132 national research organizations v.s. the affiliated researchers' population in the following figure.

Fig 1 shows that larger organizations (larger than 2,000 approximately) have relatively small variation in the numbers of researchers that can be found in different database products. On the other hand, in the case of smaller organizations ($<$ 1,000) the identified numbers of the researchers are diverse.

As described briefly, the bipartite graph edge is established on two factors: the affiliation and the researcher's name. This construction has an obvious disadvantage of multiple people with the same name. In Japan, we fortunately have relatively diverse names, which is explained by the Galton process, a stochastic process named after 19c. English researcher Francis Galton. However, we only have 151,236 names for 181,478 national organization researchers. By considering the affiliation, 4,339 researchers are left without unique addressing key. If the

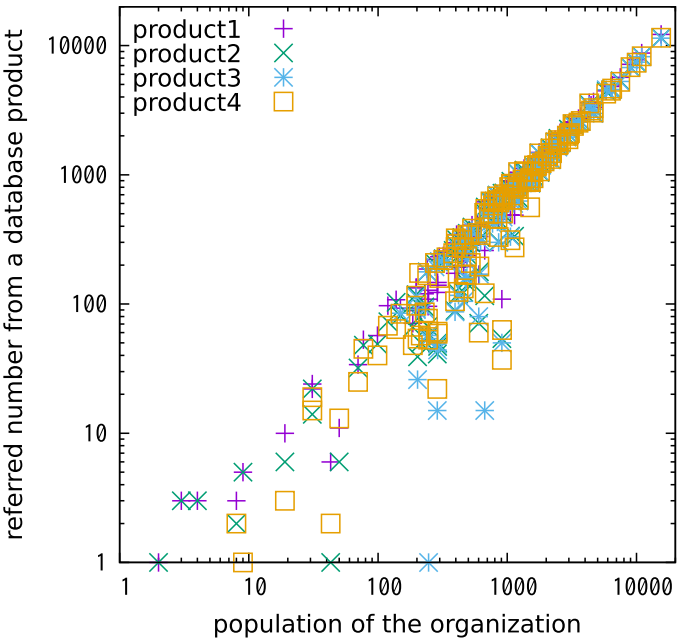

Fig 1. The general data presentation. Each dot represents the organization population (the horizontal axis) v.s. the number of researchers identified as authors in particular database product (the vertical axis). The plot exists below the diagonal line because not all the affiliated people are found in the research database.

https://doi.org/10.1371/journal.pcsy.0000020.g001

organization has a large number of researchers, then the problem of identical names is more likely to be serious.

Affiliation identification plays an important part in our author identification process. We use an identification system developed and maintained by National Institute of Science and Technology Policy (NISTEP for short), a department of the Ministry of Education, Culture, Sports, Science and Technology.

The author identification procedure is not particularly new, nor advanced. It is simple and deterministic which enables us to reproduce identical results from identical data. The procedure requires $O(n)$ order of working space and time for the population of the researchers n. $O(n)$ is not particularly scalable, yer we consider it to be acceptable. Additionally, it is simply implemented by classical information processing techniques, which ensures it to be shared among the Japanese Government members.

The model for recall estimation

There are several different research database products available in the market. We give index $i = 1...n$ to the database products (which is the authors' set) X_i, its associated bipartite graph G_i, and researchers' set (the set of referred entities) Y_i. The set Y_∞ is the truth set, the perfect set of researchers already appeared in Eq 2, which is basically unavailable because no research database product or identification procedure is perfect.

Here is the heuristics of our model of recall estimation. Suppose we have a few different research database products and associated graphs. It seems almost trivial that for $i \neq j$, Y_i and Y_j are not the same set. Actually, this inequality provides information to reach Y_∞. The difference comes from the fact that there are researchers who appear in a certain database product and are missing in other products, which means the set $Y_\infty - Y_i$ is not empty. If the set $Y_\infty -$

Y_i has only a few members, then it is less likely to find a new researcher when we obtain an alternative research database product. Conversely, if the set $Y_\infty - Y_i$ is large, it is more likely to find a new researcher. Suppose we have two database products, i, j, then the set $Y_i \cup Y_j$ will leave even fewer researchers who are to be identified. Repeating the above steps by obtaining another research database, $\cup Y_i$ will eventually reach Y_∞, with the gain getting progressively smaller.

It is not very useful to try set unification over the authors' sets because these sets are provided by different database vendors with their own identification systems. However, thanks to the e-Rad system, the set of researchers can be usefully unified.

We illustrate the concept of our model visually in Fig 2. From the researchers' set $\{Y_i\}$, $i = 1...n$, a partially ordered chain can be generated by set unification as: $Y_i \subset Y_i \cup Y_j \subset Y_i \cup Y_j \cup Y_k \ldots$. If we have n database products, we will also have n researchers' set as $\{Y_i\}$. Therefore, out of n researchers' set, we will have $n!$ such chains.

Let $(\cup_{i=1}^{1} Y_i, \cup_{i=1}^{2} Y_i \ldots, \cup_{i=1}^{k} Y_i)$ be one of those chains.

Let $P(Y)$ be the probability of an event $y \in Y$ for simplicity. Let $P(Y_{k+1} - \cup_{i=1}^{k} Y_i)$, the probability of finding a new researcher by obtaining an alternative database, which is represented by the light red area of Fig 2, be proportional to the multiplication of the size of the set $Y_\infty - \cup_{i=1}^{k} Y_i$ and Y_{k+1} as

$$P(Y_{k+1} - \cup^k Y_i) = c\ P(Y_{k+1}) P(Y_\infty - \cup^k Y_i). \tag{3}$$

Let $c \leq 1$ be a constant positive real number.

In the special case of $c = 1$, the events $y \in Y_i$ are independent of each other. However, because the database vendors are competing in the open market, their products are not independent, and the intersection part is larger than expected from the naive independence hypothesis, therefore c is generally smaller than 1.

Eq 3 means that by considering the new database product, the set $\cup^k Y_i$ will asymptotically reach the size of the set Y_∞ (hence it has the index script ∞).

The chain $\cup^k Y_i$, $k = 1...$ is a record of progressively growing author identification results as we authors (of this research) work on alternative database products. Let $\Delta P(Y_t) = P(Y_{(t+1)} - Y_t)$ for readability, suppose additionally $P(Y_i)$ is constant and assign a new constant q as $q = cP(Y_i)$.

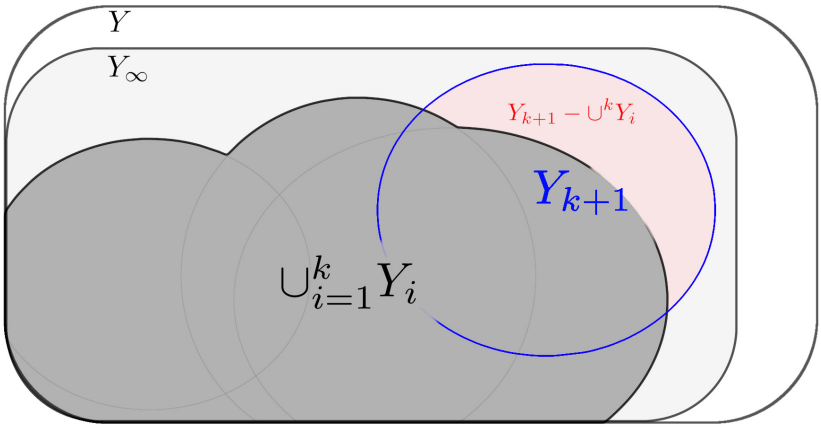

Fig 2. Visual illustration of the model. The outermost frame represents the whole researchers' set Y, within which the truth set Y_∞ resides. The last element of the set unification chain $\cup_{i=1}^{k} Y_i$ is the dark gray area. The researchers found in the new database Y_{k+1} (blue boundary with blue letters) is represented by the light red area. The newly found population is proportional to the multiply of $|Y_\infty - \cup_{i=1}^{k} Y_i|$ and $|Y_{k+1}|$.

https://doi.org/10.1371/journal.pcsy.0000020.g002

Then the expectation of ΔP can be expressed as follows:

$$E(\Delta P(Y_t)) = q(P(Y_\infty) - P(Y_t)). \tag{4}$$

Eq 4 is a discrete version of the differential equation that first appeared as the Newton's law of cooling in [1], which describes that the flow of heat is proportional to the temperature difference between the object and the environment $|P(Y_\infty) - P(Y_t)|$. Obviously, it is not limited to cooling or monotonic decreasing; in the case of this research, the value increases asymptotically to $P(Y_\infty)$.

Next, we will describe the estimation procedure based on the model of Eq 4. In order to obtain the recall value, we first have to calculate the truth set size $P(Y_\infty)$, which is to be estimated as the parameter of Eq 4. In this research, we have four database products and associated bipartite graphs, whose set of researchers Y_1, Y_2, Y_3, Y_4 as described in the data description section. Generally speaking, only four data points is not enough to estimate $P(Y_\infty)$ reliably.

Actually, we have more than four data points. The model of Eq 4 is about the *sequence* of set unification, not the vertices set itself. It describes how the element size of such set-inclusion chain increases when an alternative database product comes into consideration, where the parameter $t = 1, \ldots 4$ defines the steps of the database increment. Therefore, fitting of Eq 4 is performed to each set inclusion chain of the four sets Y_1, Y_2, Y_3, Y_4. For example, the sequence $Y_1, \cup_{i=1}^{2} Y_i, \cup_{i=1}^{3} Y_i, \cup_{i=1}^{4} Y_i$ is one of such unification sequences, and $Y_4, \cup_{i=3}^{4} Y_i, \cup_{i=2}^{4} Y_i, \cup_{i=1}^{4} Y_i$ is another one. There are total of 4! = 24 such chains.

Each set inclusion chain gives a small clue to the value $P(Y_\infty)$ because it is a realization of the model of Eq 4. Therefore, the overall likelihood of the model can be obtained as the product of the likelihood of all the set unification sequences. In other words, we look for the value $P(Y_\infty)$ that best agrees with all 24 chain variations. Based on the estimated value $P(Y_\infty)$, we can calculate the recall value of our bipartite graph.

We will prepare 24 set inclusion chains from four variations of bipartite graphs in order to estimate the $P(Y_\infty)$ value. This process is repeated over 132 national organizations.

The actual estimation is performed by the Markov-Chain Monte Carlo method. More specifically, by Hamiltonian MCMC on MC-Stan software. The method performs Bayesian inference on the probability distribution of the model based on data (observations). The nature and principle of MCMC can be found in [8]. A brief description of the algorithm is as follows:

```
(define mcmc (cfg)
  (if (or (> (lklhd (altntv cfg)) (lklhd cfg))
      (< (rand) (/ (lklhd (altntv cfg)) (lklhd cfg))))
      (altntv cfg)
    cfg))

(while (> requirement (length samples))
  (if (not (equal cfg (mcmc cfg)))
      (set 'samples (cons (set 'cfg (mcmc cfg)) samples))))
```

In the pseud-code above, the variable "cfg" is a configuration of the model, the function "lklhd" is the likelihood of the model under given configuration, and function "altntv" yields alternative configuration from given configuration. The function "mcmc" returns alternative configuration except that the likelihood of the alternative configuration is lower than the current configuration. However, even if the likelihood of the alternative configuration is

lower than the current configuration, mcmc still returns alternative configuration by the probability of

$$\frac{lklhd(altntv(cfg))}{lklhd(cfg)},$$

which is large if the alternative configuration gives similar, if not better likelihood than the current one. The algorithm stops at the current configuration (which means it does nothing) only if the alternative configuration is enough bad. Thanks to this arrangement, the function "mcmc" returns likely configuration more often than unlikely configuration, which means the collection of configurations obtained from "mcmc" function is a numerical representation of the configurations distribution of the model. We collect the samples as much as we need in the later part of the pseud-code. In this research, the optimal (the most likely) model configuration is extracted from the collected samples, along with the confidential interval.

The key to the effectiveness depends on the function "altntv", especially when the configuration has high dimension. The software utilized in this research provides an advanced way to find a better configuration by preliminary sampling and obtain an approximate idea of the model, based on which actual sampling is effectively executed. The technical background about finding the alternative configuration is described in [9], and is implemented in C++ in the software suit.

The working time and space of the model are not affected by the number of vertices (population to be processed). However, as the data preparation procedure generates Stirling number order of data sequences, it is not scalable to additional observations. When several times more observations are available, the data preparation procedure requires to be re-designed.

The data flow diagram is presented here to show the overall data preparation and processing procedure. The model predicts truth set size from multiple observations, which are Y_i, ($i = 1...4$) in the upper left corner of Fig 3. The observations are processed to set-unification

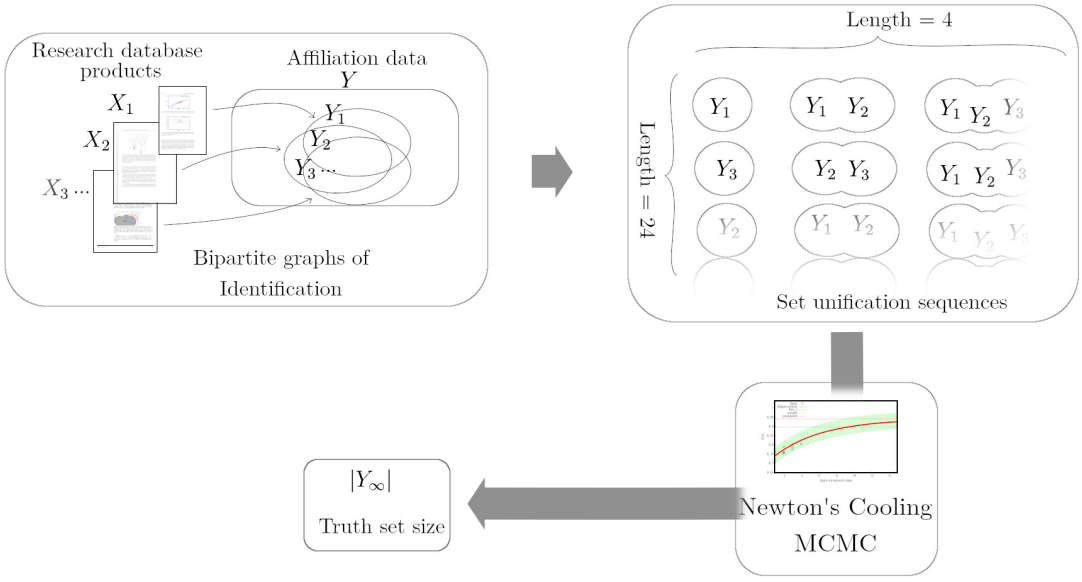

Fig 3. The overall data flow diagram of this study. The model, which is Newton's cooling law implemented by MCMC, predicts the truth set size from the set-unification sequence data. The set-unification sequences are prepared from four directed bipartite graphs of author identifications, which are constructed from four research database products and Japanese national research organizations' affiliation data.

https://doi.org/10.1371/journal.pcsy.0000020.g003

sequences and each unified set size is measured. In this study, there are 24 sequences, based on which the model predicts the truth set size.

Results

Precision

First, we calculate the precision of the obtained bipartite graph by calculating the proportion of unique addressing keys.

As described in the Data section, the construction of the bipartite graph is based on the combination of the affiliation information and the person's name (first name and family name). The entity addressed by a shared symbol is ambiguous, but a unique symbol cannot always assure the validity of the identification. However, researchers today have little motivation to fake their names nor affiliations on their academic papers. Therefore if we find an author with a certain name and affiliation, who is also found in the e-Rad system with a unique and exact name and affiliation combination, there is almost no reason to doubt the identification.

Fig 4 shows the cumulative distribution of precision of the identification result from two different point of view. The "+" symbol plot shows the distribution of organization-wise values. It shows the proportion of the 132 organizations that marked the value or lower. The achieved precision is higher than 0.98 in more than 80 percent of Japanese national research organizations. We are also concerned about how much of the total researcher is covered by the precision, which is shown in the "x" symbol plot.

If a large number of people are affiliated to a certain organization, it is more likely to find people with identical names, which lowers the precision of the bipartite graph. Therefore, the population-based "x" symbol plot lies consistently left to the "+" symbol plot, which means precision is lowered if we consider the organization size. Despite the limitation of the relatively primitive method to establish the links, we can claim that we could cover more than 80 percent of researchers by precision 0.95. The overall precision is 0.97. The achieved identification

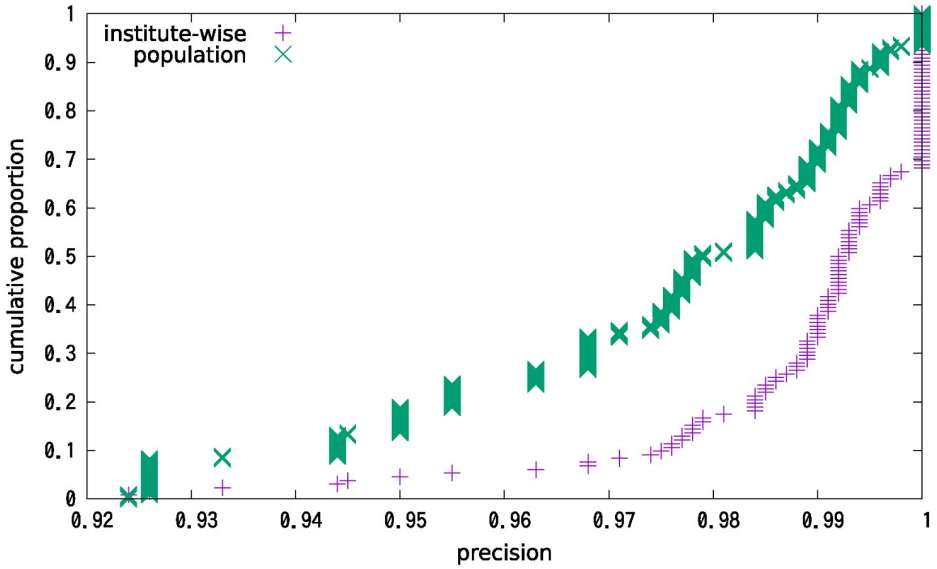

Fig 4. Precision cumulative distribution of identification result. The plot shows the proportion of the organizations (in the "+" symbol) and the proportion of the researchers (in the "x" symbol) covered by the obtained precision.

https://doi.org/10.1371/journal.pcsy.0000020.g004

seems to be good enough for macroscopic whole-nation scale analysis, only if we could obtain the recall.

Recall

In this sub-section, we will describe the recall estimation of a typical large-scale national organization. As we discussed in the Method section, we must obtain the truth set size $P(Y_\infty)$ in order to obtain the recall value. Therefore, most of the recall estimation procedure is to calculate the value of $P(Y_\infty)$ by MCMC.

Here in Fig 5, the organization we selected is Nagoya University. Nagoya University is a well-known national university located in Nagoya whose history can be traced back to the late 19c. It has departments of agriculture, economics, education, engineering, informatics, law, literature, medicine, and science to which in total approximately 3E+03 researchers belong.

Based on the estimation $P(Y_\infty) = 0.94$ and the destination set size of the constructed bipartite graph 0.76, the recall value of Nagoya University is $\frac{0.76}{0.94} = 0.81$. This value is a typical one found among large scale national organizations of Japan.

The estimation of MCMC is executed by numerical sampling, generating the distribution of the estimated values. Therefore, the confidential interval of the estimation is also obtained along with the expectation, which is plotted in the pale green band in Fig 5.

We present this particular case because we have author-researcher identification data manually and carefully prepared for the purpose of university management, and separately from this research. The data is expected to contain a maximal set of researchers who can be found in Scopus database.

According to this data, the proportion of the authors is 0.897 (the gray horizontal line), which can be regarded as an empirical limit-inf of $P(Y_\infty)$ for the following reasons:

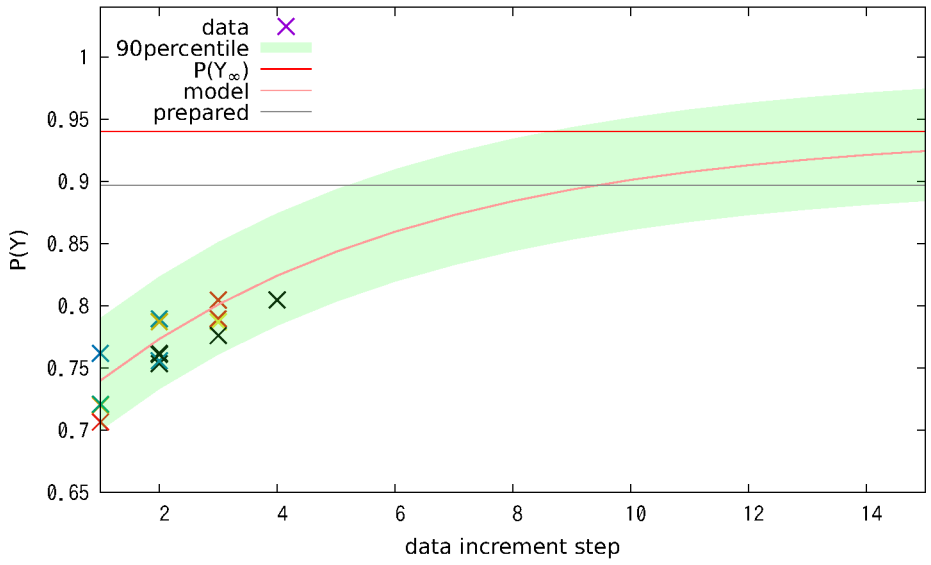

Fig 5. The $P(Y_\infty)$ estimation of Nagoya University case. The horizontal axis is for the data increment steps t of Eq 4, and the vertical axis is for $P(Y)$ value, showing that the estimated model will reach the truth set size $P(Y_\infty)$ asymptotically. The "x" symbols in the lower left corner represent the data, which are 24 set-inclusion chains of length 4, each chain with a different color(some are visually overlapped). The estimated value $P(Y_\infty) = 0.94$ is in the red horizontal line to which the model (pale red line) converges. The 90 percentile confidential interval of the estimation is in the pale green band. The gray horizontal line at 0.897 is the value from separately prepared data based on Scopus.

https://doi.org/10.1371/journal.pcsy.0000020.g005

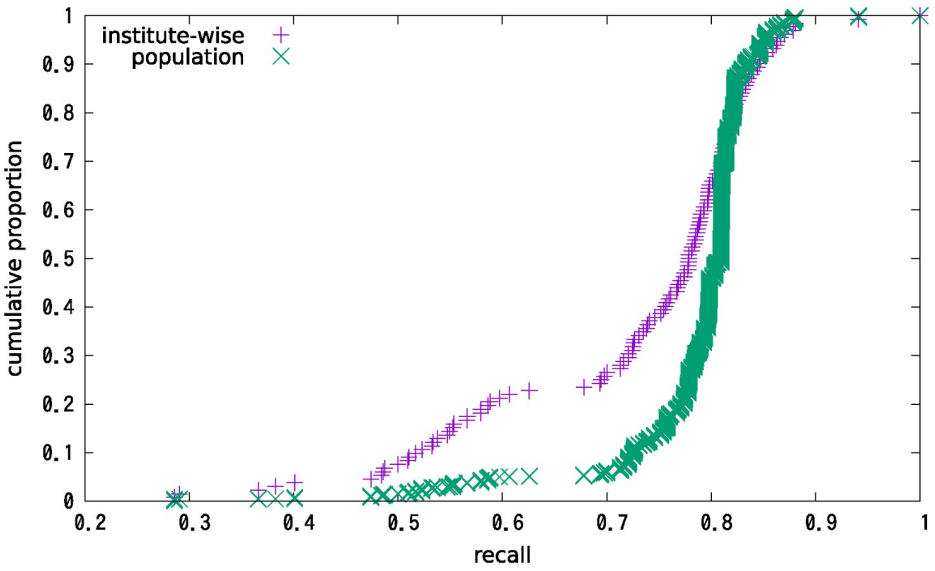

Fig 6. Estimated recall value distribution. The plot is configured identically to Fig 4 to see the distribution of institute-wise count and actual population.

https://doi.org/10.1371/journal.pcsy.0000020.g006

- As the value is based on a single database product, the value should give the lower-bound of the value $P(Y_\infty)$ because $P(Y)$ should be necessarily increased by considering alternative database products.
- Also, the value is a result of the practically best exploitation of a single research database product.

Actually, the 90 percentile confidential interval of $P(Y_\infty)$ estimation is [0.9, 0.99], which is in concert with the empirical limit-inf given above.

Fig 6 shows that 80 percent of the total population is covered by the recall value of 0.8. As we will see in the Discussion section, typical organization generally has a large number of affiliated researchers and is estimated to have recall values of approximately 0.8. In fact, non-typical organizations found in the bottom part of Fig 6 have relatively limited presence in the overall result of our recall estimation. Therefore, unlike the case of precision of the previous subsection, the population-based plot ("X" symbol) shows a better outcome than the institute-based ("+" symbol) plot. The overall recall is 0.79.

Discussion

In this section, we will examine the result of the previous section.

Precision-Recall Relation and F1 Score

Fig 7 shows the plot of F1 score, which is the harmonic mean of precision and recall, with the identical plot configuration of Figs 4 and 6. F1 score is presented here to moderate the complex relation of precision and recall and to provide smooth understanding.

The plot shows that approximately 90 percent of the population is covered by a score of 0.85. As discussed in the previous sections, relatively large organization tends to show high recall value with lower precision. The lower precision of large organizations comes from name collision, which cannot be avoided when addressing large numbers of people with their

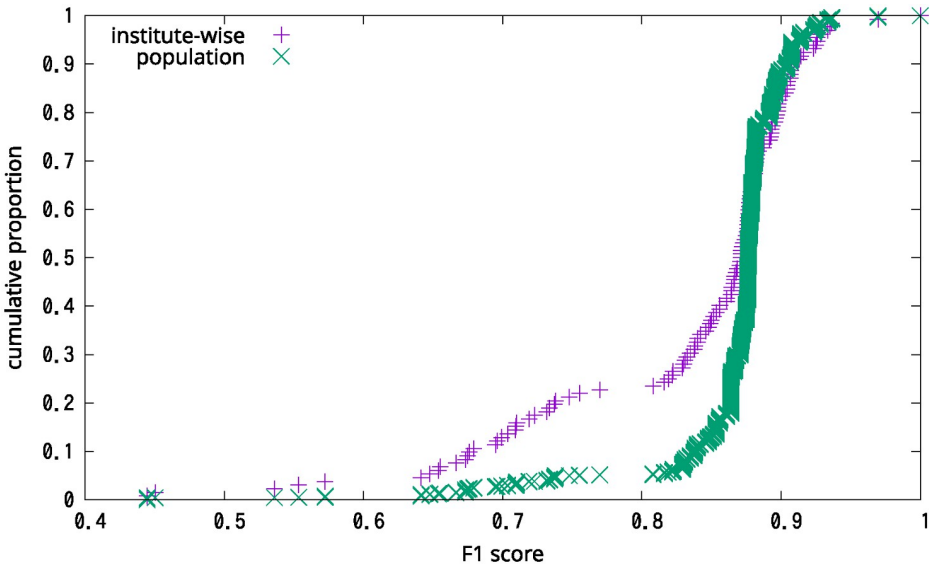

Fig 7. The distribution plot of F1 score, which is $\frac{precision * recall * 2}{precision + recall}$ with identical plot configuration to the Figs 4 and 6.

https://doi.org/10.1371/journal.pcsy.0000020.g007

names. On the other hand, major organizations tend to have a large set of consistently identified researchers, which yields high recall value, with a stable and sound estimation process. In summary, the combination of high precision and high recall is hard to come by, although trade-off of these two indices is not universally valid (see [10]).

Fig 8 shows the recall value versus the standard deviation in $P(Y_\infty)$ estimation. MCMC executes estimation by generating the sample distribution of the estimated values, hence not only the expectation but also the standard deviation of the estimation is obtained. We can see a group of typical organizations characterized by a recall of approximately 0.8 and a standard deviation of 0.02 (or less). This group of organizations includes Nagoya University of previous subsection. We also see a strong negative correlation between recall value and the standard deviation in estimation.

Apart from typical organizations like Nagoya University, there are different organizations in the upper-left corner of Fig 8. The leftmost organization, which has a moderate size of affiliated population (approximately 900), is estimated to have the value $P(Y_\infty) = 0.42$ and we could find only twelve percent of the affiliated researchers. Hence its recall is approximately 0.28 as we see in the plot.

As discussed in the model description part, the variation in the observation is proportional to the complement set size $P(Y_\infty - \cup^t Y_i)$ (light gray area of Fig 2), which has a large proportion if the observations Y_i are small. Proportionally larger $Y_\infty - \cup^t Y_i$ brings proportionally larger noise into the series $P(\cup^t Y_i)$. Because of this dominant noise, the estimated $P(Y_\infty)$ from the data of this organization is widely diverse. In fact, the set-unification sequences of this organization do not look very much like Newton's cooling. An additional parameter is needed to obtain an ergodic Markov chain, which is described in S1 Text page 2.

Robustness against Noise

As discussed in the previous subsection, the model of this research does not produce equally confident results for all 132 national organizations. Also, it is the noise in the four

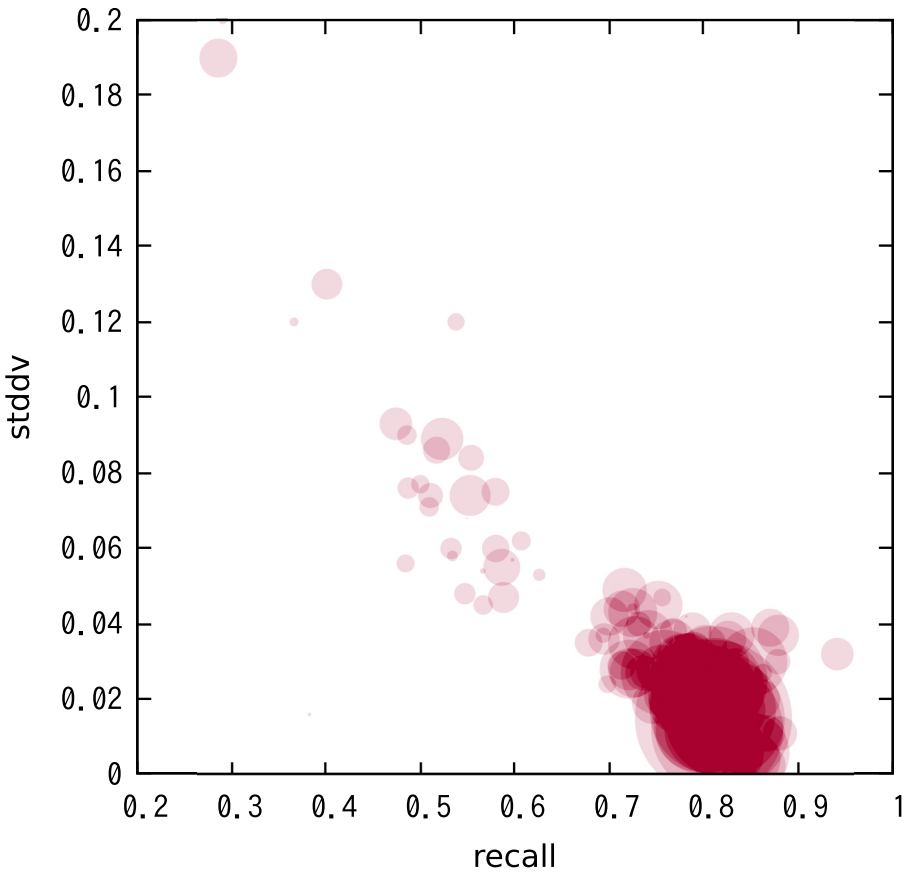

Fig 8. Recall value (horizontal axis) versus the standard deviation in the estimation (vertical). The plot is configured to have circle areas proportional to the organization populations, similar to Fig 7.

https://doi.org/10.1371/journal.pcsy.0000020.g008

identification results that enables the model to work. Therefore, It should be useful to see the impact of noise on estimation and examine the robustness of the method.

We suppose that the observation values $|Y_i|$ is subject to the binomial distribution, whose mean is equal to the observed value $|Y_i|$ with the probability $\frac{|Y_i|}{Y}$. For the distribution of the set-unification sequence increment, we also suppose binomial distribution whose mean is also the observed value $|Y_{k+1} - \cup^k Y_i|$ with the probability of $\frac{|Y_{k+1} - \cup^k Y_i|}{|Y_{k+1}|}$, which is the proportion in the newly added identification result.

After generating 100 variations from the data of Nagoya University by the procedure described above, the proposed model, Newton's cooling law on MCMC, is applied.

Fig 9 shows the estimation result distribution under the noise just described. Besides that the estimated mean is shifted upwards by 0.01, the outputs have sharp peaks that perfectly agrees with the results of Fig 5. It can be concluded that if the data looks like Newton's cooling law, the method of this study can correctly predict the truth set size against noise.

However, in the case of the organization discussed in the last paragraph of the previous subsection, the proposed model hardly converges. The prediction is already unreliable, which should be directly hit by the noise. Fortunately, such cases remains minor. Most of the population is consistently identified in multiple research databases, and the truth set size prediction model successfully converges to reliable value.

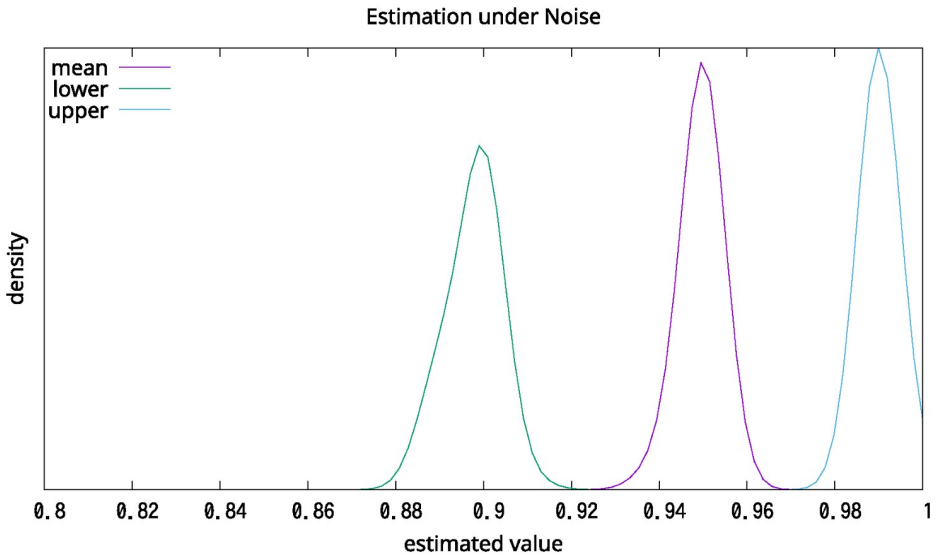

Fig 9. Distribution of the output under noise. The mean, 90 percentile upper bound, and lower bound are plotted. Unlike other distribution figures which show cumulative proportion, this plot is transformed to probability density by Gaussian kernel with the bandwidth of 0.005 to display the peak.

https://doi.org/10.1371/journal.pcsy.0000020.g009

Conclusion

Precision and recall are two useful ways to evaluate various methods, operations, and processes. However, they are often unavailable in the real world applications, especially in survey or inquiry-type process, where the truth set is basically unknown.

In this research, we developed a method to estimate the size of the truth set, which is required to obtain the recall value, by

- focusing on the *size* of the truth set (not the set itself),
- conceiving a straightforward probabilistic model,
- apply the model to the observation by way of MCMC.

The model itself is simple and classical; actually, it is one of the first mechanics formulated as a differential equation by Isaac Newton, which is known as the Newton's cooling law today.

The data, to which the model is applied, is a comprehensive nation-wide scale collection of identification results of the authors of academic article as a researchers' affiliation data of Japanese national research organizations.

For typical organizations, the estimated recall values are 0.8 or above, with a small deviation of estimation. On the other hand, in the case of some non-typical organizations, the recall values are not high, and the standard deviation of the estimation is large. However, the impact of such non-typical organizations is limited because of their small populations. Also, the model was examined against possible noise, and it turned out that the model is robust against disturbance.

In summary, most of the researchers' population is covered by our identification result with well-estimated high precision and recall value combination. In F1 score 0.85 or better result is achieved in approximately 90 percent of the total researcher population.

As the future work, we think there are two ways to proceed. One way is to improve the estimation, for example, by adding even more database products. The other way is to go beyond

the size of the vertices set and perform link prediction, for example, using information like research topics to find more researchers. The second way will also bring additional variation in the vertices set size and hopefully improve the estimation.

Supporting information

S1 Text. Truth set size prediction by Newton's cooling law.
(PDF)

Acknowledgments

The authors appreciate generous permission from Nagoya University for the use of their data of the researchers and authors relation for the university management.

The authors thank Prof. Takahiro Ueyama, full-time executive member of the Council for Science, Technology, and Innovation of the Cabinet Office for initiating and promoting of the overall project, invaluable advice and discussion for this study.

The authors also thank Mr. Naoaki Kashiwabara, director of the Cabinet Office for his direction and management of the overall project, as well as for the help to publish this study.

The authors also thank Mr. Toshihyuki Shirai, former director of the Cabinet Office for his direction and management of the overall project, as well as for the help to publish this study.

The authors also thank Mr. Iwao Miyamoto, former director of the Cabinet Office for his direction and management of the overall project, as well as useful discussion and helpful advice for this study.

The authors also thank Professor Naoshiro Shichjo of Graduate Institute for Policy Studies for his support to this research, data acquisition and invaluable advises.

The authors also thank Mr. Shinsuke Kawachi of the Cabinet Office for data acquisition and preparation from the e-Rad system.

The authors also thank NISTEP, National Institute of Science and Technology Policy, of the Ministry of Education, Culture, Sports, Science and Technology for their generous permission to use the organization identification program tools.

The authors also appreciate the invaluable advice from Ms. Maya Fujita on the visual configuration of Fig 2.

Author Contributions

Conceptualization: Yuji Fujita, Noritaka Usami, Toshiaki Fujii, Hiroaki Nagai.

Data curation: Yuji Fujita, Noritaka Usami, Toshiaki Fujii, Hiroaki Nagai.

Formal analysis: Yuji Fujita, Noritaka Usami, Toshiaki Fujii.

Investigation: Yuji Fujita, Noritaka Usami, Toshiaki Fujii, Hiroaki Nagai.

Methodology: Yuji Fujita, Noritaka Usami, Toshiaki Fujii, Hiroaki Nagai.

Project administration: Yuji Fujita, Noritaka Usami, Toshiaki Fujii, Hiroaki Nagai.

Resources: Noritaka Usami, Toshiaki Fujii, Hiroaki Nagai.

Software: Yuji Fujita, Noritaka Usami, Toshiaki Fujii.

Supervision: Noritaka Usami, Hiroaki Nagai.

Validation: Hiroaki Nagai.

Visualization: Yuji Fujita, Noritaka Usami.

Writing – original draft: Yuji Fujita, Toshiaki Fujii.

References

1. Newton I. Scala Graduum Caloris. Calorum Descriptiones & Figna. Philosophical Transactions; 22:824–829.
2. Fujita Y, Usami N. Fractal Dimension Analogous Scale-Invariant Derivative of Hirsch's index. Appl Netw Sci 7, 5. 2022. https://doi.org/10.1007/s41109-021-00443-x
3. Milojević S, Radicchi F, Walsh JP. Changing demographics of scientific careers: The rise of the temporary workforce. Proceedings of the National Academy of Sciences. 2018; 115(50):12616–12623. https://doi.org/10.1073/pnas.1800478115 PMID: 30530691
4. Bagga A, Baldwin B. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics—Volume 1. ACL'98/COLING'98. USA: Association for Computational Linguistics; 1998. p. 79–85. Available from: https://doi.org/10.3115/980845.980859.
5. Ono S, Yoshida M, Nakagawa H. NAYOSE: A System for Reference Disambiguation of Proper Nouns Appearing on Web Pages. In: Ng HT, Leong MK, Kan MY, Ji D, editors. Information Retrieval Technology. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 338–349.
6. Kurakawa K, Takeda H, Takaku M, Aizawa A, Shiozaki R, Morimoto S, et al. Researcher Name Resolver: identifier management system for Japanese researchers; 2014.
7. Haixia W, Chunyao S, Yao G, Tingjian G. Link Prediction on Complex Networks: An Experimental Survey. Data Science and Engineering. 2022; 7.
8. Diaconis P. The Markov chain Monte Carlo revolution. Bulletin of American Mathematical Society. 2009; 46:179–205. https://doi.org/10.1090/S0273-0979-08-01238-X
9. Betancourt M. A Conceptual Introduction to Hamiltonian Monte Carlo; 2017. Available from: https://arxiv.org/abs/1701.02434.
10. Cleverdon CW. Journal of Documentation. 1972; 28(3). https://doi.org/10.1108/eb026538