A Method for Eliminating Articles by Homonymous Authors from the Large Number of Articles Retrieved by Author Search

Natsuo Onodera*, Mariko Iwasawa, Nobuyuki Midorikawa, Fuyuki Yoshikane
Graduate School of Library, Information and Media Studies, University of Tsukuba, 1-2, Kasuga, Tsukuba, Ibaraki 305-8550, Japan.
E-mail: {onodera, miwasawa, midorika, fuyuki}@slis.tsukuba.ac.jp

Kou Amano
Bioresource Information Division, RIKEN BioResource Center, 3-1-1, Koyadai, Tsukuba, Ibaraki 305-0074, Japan.
E-mail: amano@brc.riken.jp

Yutaka Ootani and Tadashi Kodama
Toho University Medical Media Center, 5-21-16, Omori-Nishi, Ota-ku, Tokyo 143-8540, Japan.
E-mail: {y-ootani, kodamat}@mnc.toho-u.ac.jp

Yasuhiko Kiyama
Juntendo University Library, 2-2-26, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.
E-mail: kiyama@juntendo.ac.jp

Hiroyuki Tsunoda
Department of Culture and Language, Shokei University, 6-5-1, Nirenoki, Kumamoto 861-8538, Japan.
E-mail: tsunoda@shokei-gakuen.ac.jp

Shizuka Yamazaki
International Medical Information Center, 35, Shinanomachi, Shinjuku-ku, Tokyo 160-0016, Japan.
E-mail: yshizu@plum.ocn.ne.jp

* Correspondence author

This paper proposes a methodology which discriminates the articles by the target authors ("true" articles) from those by other homonymous authors ("false" articles). Author name searches for 2,595 "source" authors in six subject fields retrieved about 629 thousands articles. In order to extract true articles from the large amount of the retrieved articles including many false ones, two filtering stages were applied. At the first stage, any retrieved article was eliminated as false if either its affiliation addresses had little similarity to those of its source article or there was no citation relationship between the journal of the retrieved article and that of its source article. At the second stage, a sample of retrieved articles was subjected to manual judgment, and utilizing the judgment results, discrimination functions based on logistic regression were defined. These discrimination functions demonstrated both the recall ratio and the precision of about 95% and the accuracy (correct answer ratio) of 90-95%. Existence of common coauthor(s), address similarity, title words similarity and interjournal citation relationship between the retrieved and source articles were found to be the effective discrimination predictors. Whether or not the source author was from some specific countries was also one of the important predictors. Furthermore, it was shown that a retrieved article is almost certainly true if it was cited by, or cocited with, its source article. The method proposed in this study would be effective when dealing with a large number of articles whose subject fields and affiliation addresses vary widely.

Introduction

Homonymous authors (each of whom has an identical family name as well as a given name) present a significant problem in article search based on author name or analysis of the productivity of individual researchers (see Chapter 14 in Moed, 2005). The problem is even more significant in cases of using a database such as ISI Science Citation Index that represents author names only by the last names and first (and second, sometimes) name initials. For countries such as Japan, China, and South Korea, where some last names are extremely common, author search based on last name and first name initial would retrieve a large number of unwanted articles by other homonymous authors. Moed (2005) showed in p. 182 of this reference that approximately 2,100 author names appearing more than 50 times per annum were found from the articles published during the period between 1999 and 2002 in the Web of Science (WoS), and among these names 65% were Asian (54% are Japanese). Obviously, many of the author names correspond to different authors. The problem with Western author names is comparatively less critical but nevertheless exists and cannot be ignored. Aksnes (2008) showed that if the 31,135 researchers registered in the Norwegian Research Personnel Register (Ver. 2005) were listed in the ISI style, 4,362 (14%) homonymous authors would be found.

A number of methods exist to discriminate the wanted articles from the unwanted ones by homonymous authors among the articles retrieved through author search. The best method is to obtain a list of the papers published by the target researcher(s) and compare the retrieved articles with those in the list (Rinia, van Leeuwen, van Vuren, & van Raan, 1998; van Raan, 2006; Bornmann & Daniel, 2007) but this method is scarcely available, except for the case of a few target researchers of limited institution(s). Otherwise, analyzing a large number of author names and corresponding articles would require a tremendous amount of labor. A widely used method involves discrimination based on the authors' affiliations and the topics of article found in databases. However, performing this task manually for a large number of articles is impractical and does not ensure perfect results. In some cases, homonymous researchers might have the same affiliation or might work on the same topic. In other cases, researchers may change affiliations and research themes.

Some studies revealed that coauthorship information is highly effective for disambiguating the homonymous authors. Wooding, Wilcox-Jay, Lewison, & Grant (2006) proposed an algorithmic method based on using information related to the research themes and funding organizations and information on the coauthors; they reported a recall ratio (ratio of correct answers to all articles for the target author) of 99% and a precision (ratio of articles by the target author in the hit answers) of 97%. These high ratios might be the result of using the researchers in a limited discipline who received funding from a specific organization as the target.

Kang et al. (2009) examined a method clustering Korean same-named authors appearing in IT-related conference papers based on their coauthors. They enhanced the coauthorship information by including 'implicit' coauthors obtained from the Web in addition to 'explicit' coauthors known from the target papers, and could disambiguate the homonymous authors with a recall ratio of 87% and a precision of 88%.

Some groups have developed author disambiguation methodologies by clustering bibliographic records including a same author name (generally last name and first initial) in common into sets of the records corresponding to individual authors. The variants for clustering are metadata features included in the records and (in some cases) information collected from other sources such the Web.

Giles and his colleagues proposed two supervised approaches - the naïve Bayes model and the support vector machine (SVM) model (Han et al. 2004) –and three unsupervised approaches – the K-way spectral clustering model (Han, Zha & Giles 2005), the SVM-DBSCAN model (Huang, Ertekin & Giles 2006) and the topic-based model (Song et al. 2007). The last two approaches took into account the problem of transitivity violations. From the results of clustering records of a same author name sampled from DBLP (Digital Bibliography & Library Project) and CiteSeer, they reported the topic-based model, which associates authors and documents through the author-topic and document-topic relations, performed the best in their supervised and unsupervised approaches. Although main attention of Giles' group seems to be focused on finding a superior clustering algorithm rather than seeking for features of higher discriminating ability, they showed that coauthor information is the most discriminating feature if available and article title is better feature than journal title for author disambiguation.

Recently, Cota et al. (2010) presented an unsupervised heuristic-based hierarchical clustering method, in which aggregated information about fused clusters is used for the next round of fusion. They claimed that this method, using the discriminating attributes same as those used by Giles' group (coauthor names, article title and journal title), performs comparably with, or better than, the supervised (Han et al. 2004) and unsupervised (Han, Zha & Giles 2005; Huang, Ertekin & Giles 2006) methods above-mentioned.

The group of McCallum has investigated into representations that enable 'aggregate' (triplet or higher order) comparison among bibliographic records for author disambiguation, in which information beyond pairwise comparisons is available and the problem of transitivity violations can be avoided (McCallum & Wellner 2003; Kanani, McCallum & Pal 2007). By applying this method to partitioning of a graph whose nodes are DBLP sample records with a common author name and whose edges represent feature similarities between linking nodes, they confirmed that the clustering performance was remarkably improved either by increasing the weight on an edge if hits were obtained from a Web query concatenating titles of the two nodes linked by the edge or by adding the Web page retrieved by the query as a new node of the graph (Kanani, McCallum & Pal 2007; Kanani & McCallum 2007).

Torvik and his group proposed a sophisticated algorithm which predicts the probability that a pair of MEDLINE records including a same author name in common is authored by a same person, based on a similarity profile between the two records (Torvik et al. 2005; Torvik & Smalheiser 2008). From an large-scale experiment using the 2002 baseline version of MEDLINE (Torvik et al. 2005), they showed the most important feature, among those included the similarity profile, for disambiguation of authors was the existence of common coauthor(s), followed by being published in a same journal and the agreement of middle name initial, and reported a high performance (a recall ratio of 91.9% and a precision of 98.5%) of this methodology. Torvik & Smalheiser (2008) applied the enhanced model to the 2006 baseline MEDLINE version with assessment of the model from various aspects. They revealed that addition of first full name and e-mail address (these data were collected not only from the MEDLINE records but also from the Web) as disambiguation features is effective for attainment of an excellent performance of the enhanced model, together with other improvements including correction of transitivity violations.

The ASE (Approximate Structural Equivalence) algorithm recently proposed by Tang & Walsh (2010) is unique in that information on references commonly cited by two articles is used as the similarity measure of the article pair, with a higher weight assigned to a less cited reference. This method was proved to be very effective for records having reference(s) common to other record(s), but records having no common reference are regarded as singletons since Tang & Walsh did not use other discriminating attributes.

In a recent comprehensive review, Smalheiser and Torvik (2009) reported numerous methods of author name disambiguation examined in many studies, including those not mentioned above.

Objective

Within the framework of another study that we have conducted,[1] we encountered a problem to extract only articles truly written by target authors from a large number of retrieved articles (hereinafter referred to "retrieved articles") through search by author name of specified articles (referred to "source articles"). Author name searches (by last name and first name initial) of the WoS on approximately 2,500 "source" authors provided a total number of over 600,000 retrieved articles. Separating these articles into ones that were authored by the source authors and ones that were not (hereinafter referred to as "true articles" and "false articles," respectively) cannot be accomplished manually. Thus, we examined a semiautomatic method to discriminate between true and false articles. In this paper, we propose the methodology of this approach and demonstrate its effectiveness.

As described in the Section "Background", the approaches used by Kang et al. (2009), Giles group (Han et al. 2004; Han, Zha & Giles 2005; Huang, Ertekin & Giles 2006; Song et al. 2007), McCallum group (McCallum & Wellner 2003; Kanani, McCallum & Pal 2007; Kanani & McCallum 2007) and Torvik group (Torvik et al. 2005; Torvik & Smalheiser 2008) aim to partition bibliographic records by some different homonymous authors into clusters comprising records of individual authors. For this purpose, similarities between records for all pairs in the set have to be calculated in principle. The purpose of our study is, on the other hand, to discriminate whether a given article retrieved by author name search is one by the source author or not, so it is not necessary to compare similarities among all retrieved articles (although there may be the cases that comparison between retrieved articles helps discrimination of these articles). For this reason, we adopted an approach comparing each retrieved article with its source article, not clustering a retrieved article set. In this sense the aim of our study resembles that of Wooding et al. (2006), but our study deals with a much larger article set whose authors are from more diverse subject areas and affiliated organs.

We decided to utilize as numerous features obtainable from the WoS database as possible for discriminating true articles from false ones. They are as follows:

- coauthor information of source and retrieved articles

- affiliation addresses of source and retrieved articles
- citation relationships between the journals of source and retrieved articles
- title words of source and retrieved articles
- interval between the years of publication of source and retrieved articles
- the source author’s affiliation country
- citation and cocitation relationships between source and retrieved articles

We aim to achieve a correct answer ratio of at least 90%, knowing some limitations of our approach such as use of information obtainable only from the WoS database as the discriminating features and not taking the problem of transitivity violations into account.

It should be noted that the problem of the same person having different names (because of family name changes, etc.) is outside the scope of this study.

Set of Articles Subjected to Author Discrimination

Our source articles were sampled out from normal articles published in the year 2000 in 24 journals belonging to the following six subject fields.

- <u>Condensed matter</u> physics
- <u>Inorganic</u> and nuclear chemistry
- <u>Electric</u> and electronic engineering
- <u>Biochemistry</u> and molecular biology
- <u>Physiology</u>
- <u>Gastroenterology</u>

Hereafter, these fields will be referred to using the terms underlined above. The abovementioned fields were selected because they are narrower fields that represent the broader fields of physics, chemistry, engineering, biology, experimental medicine, and clinical medicine.

About 60 source articles from each journal were randomly sampled out, amounting 1,395 as total. We selected as the source authors all authors of the source articles in 6 journals (one per field) and only the first authors of the source articles in the remaining 18 journals. Thus, 2,595 source authors were selected. Table 1 shows the 24 source journals together with the numbers of the source articles and source authors allocated to these journals.

We conducted an author name search for each of the source authors using the WoS database. The search was done for articles published till 2000, the year of publication of the source articles. Since the WoS database we used allows retrospective search of articles published since 1970, we expect to cover the entire active publishing period of almost all the source authors.

Table 1. Journals from which the source articles are extracted.

Field		Journal	Source articles	Source authors
Condensed matter		European Physical Journal B	60	60
		Journal of Physics - Condesed Matter	56	56
		Physica B	59	59
	*	Physical Review B	55	182
Inorganic		Inorganic Chemistry	53	53
		Inorganica Chimica Acta	60	60
	*	Journal of the Chemical Society - Dalton Transactions	54	249
		Transition Metal Chemistry	60	60
Electric		IEE Proceedings - Circuits, Devices and Systems	51	51
		IEEE Transactions on Circuits and Systems I - Fundamental Theories and Applications	60	60
	*	IEEE Transactions on Microwave Theory and Techniques	59	209
		Signal Processing	59	59
Biochemistry		European Journal of Biochemistry	60	60
		Journal of Biochemistry (Tokyo)	60	60
	*	Journal of Biological Chemistry	60	296
		Journal of Molecular Biology	60	60
Physiology		Japanese Journal of Physiology	60	60
		Journal of General Physiology	60	60
	*	Journal of Physiology - London	58	222
		Pflugers Archive.	58	58
Gastroenterology		American Journal of Gastroenterology	58	58
	*	Gastroenterology	59	387
		Gut	56	56
		Journal of Gastroenterology	60	60
		Total	1395	2595

All authors are selected from journals with * and only first authors otherwise.

Almost all author names in the WoS are represented by last name and first name initial (e.g., "Smith, A") or by last name and first and second name initials (e.g., "Brown, AB"). We used the spellings of author names same as described in the source articles as the searched author names.

A total of about 629 thousands retrieved articles (excluding the source articles themselves) were obtained by querying for the 2,595 source authors. That is, the number of retrieved articles per source author is 242. This number is much larger than the average productivity of a researcher during at most 30 years (1970-2000). This clearly indicates the existence of a large number of homonymous authors corresponding to the source authors.

Method for Discriminating True and False Articles

In this section, we discuss the procedure used for classifying the retrieved articles into true and false articles. The outline of the whole procedure is shown in Figure 1.

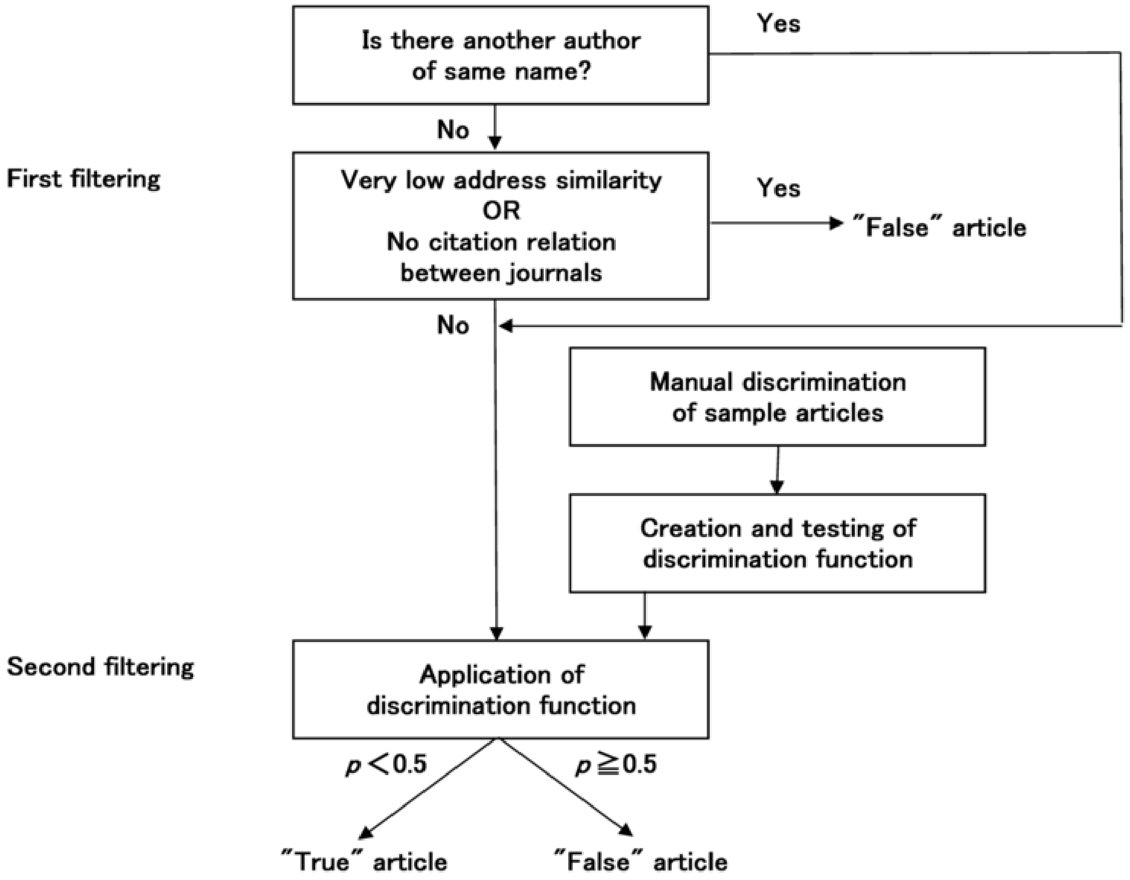

Is there another author of same name?
Yes
No
First filtering
Very low address similarity OR No citation relation between journals
Yes
"False" article
No
Manual discrimination of sample articles
Creation and testing of discrimination function
Second filtering
Application of discrimination function
p<0.5
p≧0.5
"True" article
"False" article

Figure 1. Outline of the true-false article discrimination procedure.

Information Used for Discrimination

(1) Coauthors of source and retrieved articles

If at least one coauthor name other than the source author in a retrieved article matches with one in its source article, the probability that the retrieved article is a true article will be high.

(2) Affiliation addresses of source and retrieved articles

Since the notations of the addresses of authors' affiliations vary significantly, the probability of obtaining an exact match between the addresses provided in source and retrieved articles would not be high. However, it is obvious that the higher the degree of similarity between the notations used in a retrieved article and its source article, the higher will be the probability of the retrieved article being true.

(3) Citation relationships between the journals of source and retrieved articles

If the journal in which a retrieved article was published has very little or no citation relationship (neither citing nor cited by) with the journal in which its source article was published, the probability of the retrieved article is false will be high.

(4) Title words of source and retrieved articles

Since articles written by a same author are likely to have common words in their title compared to ones by different authors, the weighted similarity of title words between source and retrieved articles would be one of the effective discriminating factors. Abstract words are not used here since the words in abstracts are very diverse and also there are a considerable number of WoS records which do not contain abstract.

(5) Citation of retrieved articles by source articles

If a retrieved article is cited by its source article, the probability of the retrieved article being true would be considerably high.

(6) Cocitation between source and retrieved articles

If a retrieved article and its source article are cited by at least one common document, i.e., the two articles are cocited, it is likely that the two articles are written by a same author.

Bibliographic coupling between a retrieved article and its source article (sharing of at least one common reference by the two articles) would also give important information for author disambiguation, as shown by Tang & Walsh (2010). But we did not use this feature since it was substantially impossible for us to check the all references of about a hundred thousands of retrieved articles with the all references of their source articles.

(7) Interval between the years of publication of source and retrieved articles

Even if the address provided in a retrieved article differs from that provided in its source article or even if there is little citation relationship between the journals of a retrieved article and its source article, the possibility exists that the two articles are by a same author who might have changed his/her affiliation or subject field. This possibility would become higher when the publishing date of the retrieved article is considerably earlier than that of the source article. Therefore, the difference in the publication years of retrieved and source articles can be considered as a discrimination factor.

(8) Whether the source author's affiliation country is the specified one or not

A rough examination of retrieved articles revealed that homonyms are particularly frequent in China, Japan, South Korea, and Taiwan. Therefore, if the source author is from one of these four countries, the probability that its retrieved articles are false would be become high.

Pre-processing the Data for First Filtering Stage

In this subsection, we describe the method of quantifying five out of the eight features mentioned in the previous subsection. These feature values are assigned to all of the 629 thousands retrieved articles described in the Section "Set of Articles Subjected to Author Discrimination". Quantifying other three features will be explained in the Subsection "Pre-processing the Data for Second Filtering Stage".

(1) Common coauthor(s) between source and retrieved articles

The coauthor names other than the source author associated with each retrieved article are checked against those of its source article. When at least one coauthor name by the WoS description (last name and first name initial or last name and first and second name initials) matches, the *AutMatch* value for such retrieved article is set as "1" and the value "0" is assigned otherwise.

(2) Similarity of affiliation addresses of source and retrieved articles

(a) Weighting of words that appear in affiliation addresses

Data on the affiliation addresses of all the authors are extracted from all of the retrieved and source articles. The frequency distribution of the words used in the addresses (except those words corresponding to country name) is obtained. The total word occurrences in the affiliation addresses is 11.64 million, and the number of different words is 92,225. The following weights are given to these words on the basis of their frequency.

Frequency	Weight	Number of different words
100000 or more	1	16
From 10000 to 99999	2	126
From 1000 to 9999	3	1085
From 100 to 999	4	4291
From 10 to 99	5	15163
From 1 to 9	6	71544

(b) Assigning an address similarity measure (*Add_Sim*) to each retrieved article

The affiliation address(es) provided in each of the retrieved articles are crosschecked against those provided in its source article. In the case of the presence of multiple addresses in either the source or the retrieved article, all the address pairs are compared. The crosscheck procedure is described below.

1) The country in the address in a retrieved article is compared to that in its source article. The following steps are executed only if the countries are identical. If there is no same country in the addresses in the source and retrieved articles, the value of *Add_Sim* is set as zero.

2) All words excluding the part corresponding to "country" are extracted from the addresses provided in the source and retrieved articles. Words consisting only of numbers are ignored.

3) All words that are common to both the addresses are extracted. This is done regardless of the position of the word in the addresses.

4) The value of *Add_Sim* is set at the sum of the weights given these words according to the abovementioned way. In the case of multiple address pairs, the values of the summed weights are compared, and the largest value is assigned to *Add_Sim*.

Here, “GERMANY,” “FED REP GER,” “GER DEM REP,” and “WEST GERMANY” are considered to be the same country. The same applies to “RUSSIA” and “USSR.” It should be noted that the countries are identified by taking into consideration the notations used in WoS. For example, some addresses from the United States of America do not contain the word “USA” but terminate with the name of the state. Furthermore, the “country” field for United Kingdom has four variations “ENGLAND,” “SCOTLAND,” “WALES,” and “NORTH IRELAND”, which are regarded as the same country UK.

(3) Journal citation relationships

The citation relationship between the journal in which a retrieved article was published and that in which its source article was published is examined using Journal Citation Reports (JCR), Science Edition, 2004. The journals are compared on the basis of the journal name and ISSN. Both of the cited times and the citing times between the journal of the source article and that of its retrieved article during the five-year period from 2000 to 2004 are counted; and the “strength of citation relationship between journals” (X) of a retrieved article is defined as the average of these cited and citing times.

(4) Interval between the years of publication of source and retrieved articles

This feature is expressed as a quantity *Age*. Since the publication year of all the source articles is 2000, if y is the publication year of a retrieved article, $Age = 2000 - y$ for the retrieved article.

(5) Source authors’ affiliation country

The quantity *FEA* for a retrieved article is set as “1” if the affiliation country of the source author is either of four Far East Asian countries (China, Japan, South Korea or Taiwan), and as “0” otherwise.

First Filtering Stage

At the first stage, the following elimination procedure for “false” articles is carried out from all the retrieved articles, since the amount of the original retrieved articles (629 thousands) is too large to process in the next stage including logistic regression analysis. We focus, at this stage, on obtaining a set of “true-like” articles of a more manageable amount, even if some true articles might be lost to a certain degree.

1) Retrieved articles whose *Add_Sim* value as defined in (2) of the previous subsection is less than 5 are eliminated.
2) Retrieved articles with values of X, as defined in (3) of the previous subsection, equal to zero (i.e., the journals did not cite each other at all during five years) are eliminated. However, this process is not applied to the field “Electric,” in which the rate of retrieved articles that cited articles published in other journals is clearly lower compared to the

other fields.

The above processes 1) and 2) are not applied to retrieved articles with *AutMatch* = 1, as defined in (1) of the previous subsection, since these articles are likely to be "true" even if *Add_Sim* or *X* values are low.

Pre-processing the Data for Second Filtering Stage

In this subsection, we describe the method of quantifying three features mentioned in the Subsection "Information Used for Discrimination" but not explained in the Subsection "Pre-processing the Data for First Filtering Stage". These feature values are assigned to only the retrieved articles which passed the first filtering stage abovementioned.

(1) Similarity of title words between source and retrieved articles

(a) Weighting of words that appear in article titles

The weight of a word is defined based on the inverse document frequency (idf) with which the word appears in titles of source and retrieved articles. The articles included in the corpus to determine the word frequencies are selected for each of the six fields, as follows.

1) all the source articles (about 240 for each field)
2) about 2,000 retrieved articles for each field, which are randomly sampled out from retrieved articles predicted as "true" by a preliminary logistic regression applied to those having passed the first filtering stage, using the independent variables *Add_Sim*, log (*X*+1), *Age* and *FEA*.

From the titles of articles comprised of 1) and 2), words are extracted according to the following steps:

(i) split the titles into character strings with space, hyphen, comma, semicolon, colon, and left and right parentheses as delimiters,

(ii) eliminate strings not containing alphabetical characters,

(iii) eliminate stopwords, which are 22 words including commonly used prepositions, conjunctions, articles, and very general words (such as "study" and "using"),

(iv) truncate strings longer than 6 characters up to the left 6 characters,

(v) when two or more same strings exist in a title, leave only one.

Strings which are left after those steps are defined "words" here, and the weight $w_i(k)$ of a word i in field k is defined as follows.

$$w_i(k) = \log [N(k)/df_i(k)]$$

Here, $df_i(k)$ is the document frequency of the word i in the field k and $N(k)$ is the total number of articles in the field k in the corpus.

(b) Assigning an title similarity measure (*Tit_Sim*) to each retrieved article

Words are extracted from the title of each retrieved article according to the same way as described above for the titles in the corpus. These title words of a retrieved article are compared to those extracted from its source article and the sum of the weights $w_i(k)$ of matched words is set as the *Tit_Sim* value for the retrieved article. Since the corpus includes all source articles as described above, the matched words necessarily exist in the corpus.

(2) Citation of retrieved articles by source articles

We obtained WoS data of all references of all the source articles from Thomson Scientific (at present Thomson Reuters Scientific). With these data, we checked whether or not each retrieved article having passed the first filtering stage was cited by its source article. Since the record IDs and the reference IDs of the WoS database are assigned by different systems, we are not able to use them for matching. Therefore, a retrieved article and a source reference are taken as identical when all of abbreviated journal name, publication year, volume and first page of the both are same. The quantity *Cited* is set as "1" when a retrieved article is cited by its source article and as "0" otherwise.

(3) Cocitation between retrieved and source articles

From Thomson Scientific we obtained WoS data of articles citing the source articles and also of those citing the retrieved articles having passed the first filtering stage. These include all citing articles till the publication year 2006. If the record ID of at least one article citing a retrieved article is identical with that of one of articles citing the source article of the retrieved article, the retrieved article is regarded as cocited with the source article. In such cases, the quantitycocit of the retrieved article is set as "1", and as "0" otherwise.

Second Filtering Stage

The retrieved articles that passed the first filtering stage are subjected to second filtering, as described below, to find further false articles.

1) A sample is extracted from the retrieved articles and the articles in the sample (hereafter called "sample articles") are manually judged as true or false.
2) A discrimination function based on logistic regression is defined for each field using the sample articles.
3) All retrieved articles are discriminated by applying the discrimination functions defined from the sample.

(1) Manual judgment of sample articles

Five hundreds of retrieved articles from each of the six fields (i.e., a total of 3,000 articles) were sampled out. Of the 500 sample articles in each field, 400 are sampled from retrieved articles with *AutMatch* = 0 and 100 from those with *AutMatch* = 1. Under this condition, the sampling was made so that the distributions of *Add_Sim*, *X*, *Age* and *FEA* might not largely deviate from those in the population.

The ten authors of this paper (all of them are faculty members or graduate students of a library and information science school) shared the judgment task for discrimination. Each sample article was decided whether it is true or false by two judges. The judges compared the sample and source articles regarding their themes (from the titles or abstracts) and the affiliation addresses. Information on coauthorship, publication year, and the citation relationship between the journals was also considered. When needed, the original (full text) article was referred. Web-based search was also conducted. When the decision between the two judges conflicted (as in the case of 11.2% of the sample articles), a final decision was made on the basis of inspection by another judge.

(2) Modeling discrimination functions based on logistic regression

On the basis of the results of manual judgment of the sample articles, a logistic multiple regression model is developed in order to predict the probability *p* that the retrieved article is a false article. The regression is configured differently for the six subject fields considered.

Observed values of the dependent variable *Judge* for the logistic regression are “0” or “1” according to whether the result of manual judgment is true or false. The following eight independent variables are considered as the predictors for modeling, based on the discriminating features described in the Subsections “Pre-processing the Data for First Filtering Stage” and “Pre-processing the Data for Second Filtering Stage”.

1) Existence of common coauthor(s) with same name (*AutMatch*)

2) Affiliation address similarity (*Add_Sim*): When the value of *Add_Sim* defined in the Subsection “Pre-processing the Data for First Filtering Stage” is larger than 20, it is replaced with 20 because, in practice, any value of *Add_Sim* above 20 is regarded as a perfect match. Note that the range of *Add_Sim* values for the sample articles with *AutMatch* = 0 was $5 \leq$ *Add_Sim* ≤ 20 since retrieved articles for which *Add_Sim* < 5 were eliminated by the first filtering.

3) Strength of citation relationship (log (X+1)): The distribution of X exhibits a strong skewness; therefore, its logarithm (log (X+1)) is used for modeling. It should be noted that there is no sample article with $X = 0$ in those with *AutMatch* = 0 except the field “Electric”, because they were eliminated by the first filtering.

4) Difference of the publication years between retrieved and source articles (*Age*): When the value of *Age* by definition in the Subsection “Pre-processing the Data for First Filtering Stage” is larger than 20, it is replaced with 20 at regression analysis, since the number of

retrieved articles whose *Age* exceeded 20 is less than 10% of all the retrieved articles.

5) Title word similarity (*Tit_Sim*)
6) The affiliation country of the source author being either of four Far East Asian countries (*FEA*)
7) Citation by source article (*Cited*)
8) Cocitation with source article (*Cocit*)

First, we carried out logistic regression using the above eight independent variables. However, the results showed that the regression coefficients of *Cited* andcocit were not significant for all subject fields and, in addition, that the maximum likelihood estimators for those coefficients were not obtainable for most fields. This is because the *Judge* value for sample articles is almost always “0” if either of the values of *Cited* or *Cocit* is “1”. That is, paradoxically, these two predictors are too effective for author discrimination to use for the independent variables of logistic regression.

For this reason, true/false discrimination for the sample articles is carried out by the following two steps:

1) The sample articles with *Cited* = 1 or *Coci*t = 1 are discriminated as “true” in spite of the values of other predictors.
2) For other sample articles (*Cited* = 0 and *Cocit* = 0) logistic regression using six independent variables as the predictors is performed.

The regression model is as shown below.

$$\ln\,[p/(1-p)] = \beta_0 + \beta_1 \times AutMatch + \beta_2 \times Add\_Sim + \beta_3 \times \log\,(X+1) \\ + \beta_4 \times Age + \beta_5 \times Tit\_Sim + \beta_6 \times FEA \qquad (1)$$

Here, p is the probability that the sample article is a false article.

Regression analysis is conducted using SPSS v.18.0. Variable selection is not performed.

For regression analysis, sample articles of each field are divided randomly into a training set (approximately 70%) and a testing set (approximately 30%). Using the regression model obtained with the training set, tests are performed to verify whether approximately the same performances are obtained for the training and test sets. Then, using all sample articles (except those with *Cited* = 1 or *Cocit* = 1), a final regression model is defined for each field.

(3) True/false discrimination of all retrieved articles

First, all retrieved articles with *Cited* = 1 or *Cocit* = 1 are discriminated as “true”. Next, using the regression models determined, true/false prediction of all other retrieved articles that passed first filtering is carried out. The discrimination boundary is set at p = 0.5.

Results and Discussion

First Filtering

Of the 629 thousands retrieved articles, 106,163 articles (i.e., 40.9 articles per source author) passed the first filtering stage. This corresponds to 16.9% of the total number of initial retrieved articles. Of the 106,163 articles, 39,239 (37%) were articles with *AutMatch* = 1.

Discrimination of Sample Articles for Second Filtering

(1) Manual judgment of sample articles

The results of manual judgment are shown in Table 2. Of all the sample articles, 75% were judged as true articles. As can be seen from the table, most of the sample articles with *AutMatch* = 1 were judged as true. For the sample articles with *AutMatch* = 0, on the other hand, the true article ratio shows a large variation depending on the field, from more than 85% in the field "Condensed matter" and "Inorganic" to less than 50% in the field "Biochemistry". This was mostly due to differences in the distributions of the discriminating features, especially of *FEA*, among fields. Comparisons among fields under the approximately same condition of the feature values revealed minor differences in the true/false article ratio.

Table 2. The results of true/false judgment of sample articles by manual inspection.

Field	AutMatch	#Sample articles	#Articles judged as True	True article ratio (%)
Condensed matter	0	400	350	87.5
	1	100	99	99.0
	Subtotal	500	449	89.8
Inorganic	0	400	345	86.3
	1	100	100	100.0
	Subtotal	500	445	89.0
Electric	0	400	259	64.8
	1	100	86	86.0
	Subtotal	500	345	69.0
Biochemistry	0	400	190	47.5
	1	100	97	97.0
	Subtotal	500	287	57.4
Physiology	0	400	273	68.3
	1	100	99	99.0
	Subtotal	500	372	74.4
Gastroenterology	0	400	249	62.3
	1	100	98	98.0
	Subtotal	500	347	69.4
All fields	0	2400	1666	69.4
	1	600	579	96.5
	Subtotal	3000	2245	74.8

(2) Relationships between the individual variables used as predictors and the results of judgment

As a preliminary investigation of the predicting power of the independent variables, the relationships between each independent variable and the results of manual judgment were investigated. Since most of the sample articles with *AutMatch* = 1 were judged as true, these analyses were made for only those with *AutMatch* = 0.

Figures 2, 3 and 4 show the change in the true article ratio with the change in *Add_Sim*, log (*X*+1) and *Tit_Sim*, respectively.

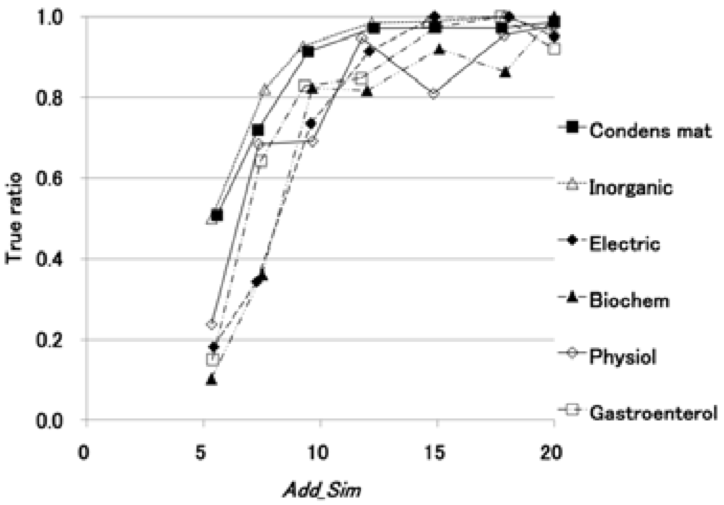


Figure 2. The relationship between the true article ratio and the address similarity (*Add_Sim*).

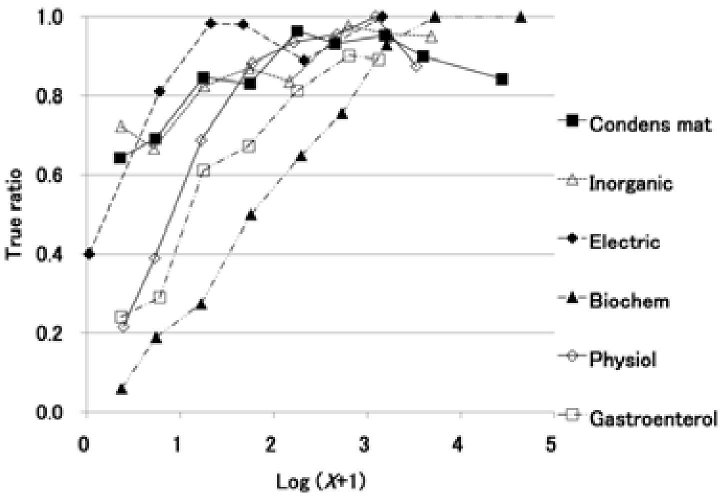


Figure 3. The relationship between the true article ratio and the strength of citation relationship [log (*X*+1)].

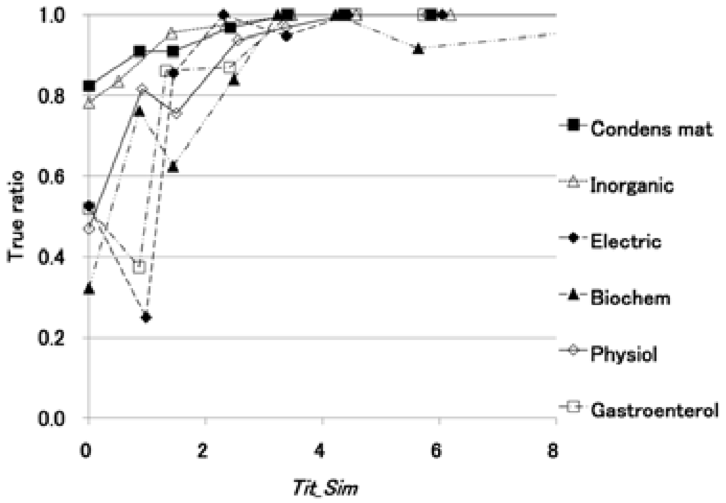
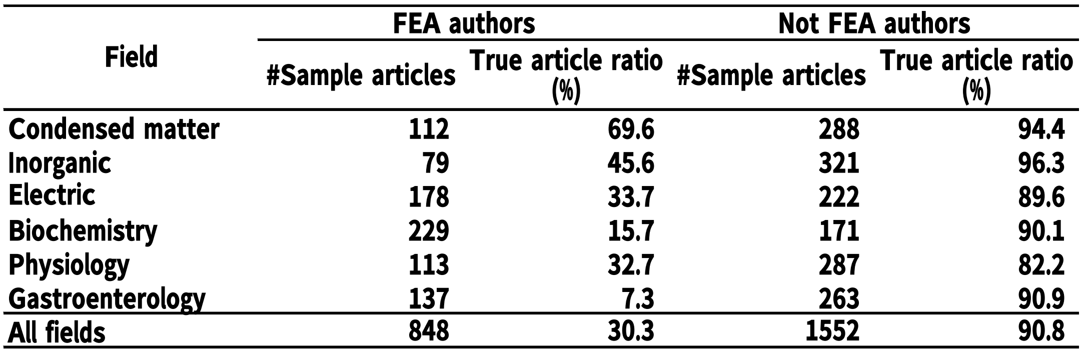


Figure 4. The relationship between the true article ratio and the title words similarity (*Tit_Sim*).

These figures show that when the value of any of the three variables increases, the true article ratio also increases. If the value of *Add_Sim* is larger than 15 or the value of Tit_Sim is larger than 4, the article is very likely to be true. The ripple observed in some bins in these graphs indicates that these bins contain fewer data points.

Table 3 shows a comparison of the true article ratios between articles for the source authors from Japan, China, South Korea, and Taiwan (*FEA* = 1) and those from other countries (*FEA* = 0). The true article ratio is 90% or more for most fields in the case of the latter group while it is considerably lower in the case of the former group. This indicates that the *FEA* is a significant predictor.

Table 3. The true article ratios for source authors from *FEA* countries in *AutMatch* = 1 articles.



Field	FEA authors		Not FEA authors	
	#Sample articles	True article ratio (%)	#Sample articles	True article ratio (%)
Condensed matter	112	69.6	288	94.4
Inorganic	79	45.6	321	96.3
Electric	178	33.7	222	89.6
Biochemistry	229	15.7	171	90.1
Physiology	113	32.7	287	82.2
Gastroenterology	137	7.3	263	90.9
All fields	848	30.3	1552	90.8

The variable *Age* has almost no correlation with the true article ratio.

Every sample article with the value of *Cited* = 1, or *Cocit* =1 was judged as true as described in the Section “Method for Discriminating True and False Articles”. (See (2) in the Subsection “Second Filtering Stage”.)

(3) Correlation among independent variables

In order to examine the possibility of the problem of multicollinearity, the correlation coefficients among the six independent variables were obtained for each field. The strongest correlation (negative) was obtained between *Add_Sim* and *FEA*, which was in the range of –0.53 to –0.25 depending on the fields. In other words, if the source author was from Japan, China, South Korea, or Taiwan (*FEA* = 1), the address similarity between the retrieved and source articles was low. The correlation coefficient between log (*X*+1) and *FEA* and that between *Tit_Sim* and *FEA* were in the range of –0.5 to –0.25 for five fields. Furthermore, the correlation coefficient between *Tit_Sim* and *AutMatch* was in the range of +0.34 to +0.15 for all fields. All other combinations of variables showed weaker or non-significant correlation. Thus, because particularly strong correlation was not observed among the variables, multicollinearity was not considered to be a significant problem.

(4) Testing logistic multiple regression models for true/false prediction

For each field, the logistic multiple regression analysis for true/false prediction was conducted based on a regression model shown by Equation (1). The sample articles with *Cited* = 1 or *Cocit* = 1 were excluded from the regression, since they were supposed to be true in spite of the values of other predictors as described in the Subsection “Second Filtering Stage” of the Section “Method for Discriminating True and False Articles”. The number of the articles with *Cited* = 1 or *Cocit* = 1 was 361 of 3,000 sample articles in all fields and ranged from 29 of the field of “Gastroenterology” to 93 of the field of “Physiology”.

The remaining sample articles were divided randomly into a training set (about 300 articles for each field) and a testing set (about 130 articles for each field) and the regression analysis was carried out for each field. The partial regression coefficients obtained are shown in Table 4 with their significance level.

The two variables *Add_Sim* and *Tit_Sim* were found to be the significant predictors for all fields. *AutMatch*, log (*X*+1) and *FEA* were also effective except for a few fields. Non-significance of *AutMatch* in the case of the field “Inorganic” was due to the fact that any sample article with *AutMatch* = 1 was not manually judged as false in this field. The variable *Age* was found to be lower in the predicting power than other variables, but still significant within the significance level of 5% for two fields. Therefore, we did not carried out variable selection and included the six variables in the regression models for all fields.

Table 4. Regression coefficients obtained from the logistic regression analysis for the training set.

Field	*n*	Const (β_0)	*AutMatch* (β_1)	*Add_Sim* (β_2)	Log (*X*+1) (β_3)	*Age* (β_4)	*Tit_Sim* (β_5)	*FEA* (β_6)
Condensed matter	311	2.40 *	-4.44 **	-0.362 **	-0.446	-0.0216	-0.778 *	1.52 **
Inorganic	319	2.61 *	-20.85	-0.386 **	-0.366	-0.0824 *	-1.368 **	2.79 **
Electric	306	3.04 **	-3.10 **	-0.374 **	-1.406 **	-0.0045	-0.946 *	2.26 **
Biochemistry	302	5.87 **	-3.88 **	-0.364 **	-1.939 **	-0.0644	-0.787 **	2.82 **
Physiology	281	3.97 **	-3.48 **	-0.251 **	-1.077 **	-0.0754 *	-0.657 **	0.75
Gastroenterology	327	2.21 *	-5.26 **	-0.247 **	-0.782 *	-0.0262	-0.653 *	3.44 **

** 1% significant, * 5% significant

When a variable *j* changes by one unit and all the other variables remain constant, it is expected that the odds ratio $p/(1 - p)$ will change by $\exp(\beta_j)$, where *p* is the false article ratio and β_j is the partial regression coefficient of variable *j*. Thus, the odds ratio decreases by 22% to 32% when *Add_Sim* increases by one unit, and decreases by 50 to 75% when *Tit_Sim* increases by one unit. Existence of common coauthor(s) other than the source author (*AutMatch* = 1) reduces the odds ratio by the factor more than 20. On the other hand, the odd ratio increases 5 to 30 times if the source author is from Japan, China, South Korea, or Taiwan (*FEA* = 1).

The true/false article prediction was performed for the sample articles of the testing set as well as the training one, using the obtained multiple regression models. An article was considered as true when the predicted value of *p* was less than 0.5 and as false otherwise. The true/false prediction boundary was tried to be set at five different values, p = 0.3, 0.4, 0.5, 0.6, and 0.7. Although the boundary value corresponding to the best performance varied between 0.4 and 0.7 depending on the fields, significant variation was not observed, so a uniform value of 0.5 was adopted. Table 5 shows the results of the prediction compared with the results of manual judgment. This table includes the results for the sample articles with *Cited* = 1 or *Cocit* = 1, all of which were discriminated as true.

From the results shown in Table 5, the following four performance measures are calculated and shown in Table 6.

Recall ratio = $a/(a+b)$

False recall ratio = $d/(c+d)$

Precision = $a/(a+c)$

Accuracy = $(a+d)/(a+b+c+d)$

Where, *a*, *b*, *c* and *d* are numbers of articles for which the manual/regression discrimination is true/true, true/false, false/true and false/false, respectively.

Table 5. The results of true/false prediction for the sample articles.

Field	Manual judgment	Prediction from the models											
		Training set			Testing set			Cited=1 or Cocit=1			Whole sample		
		True	False	Subtotal	True	False	Subtotal	True	False	Subtotal	True	False	Total
Condensed matter	True	270	6	276	112	3	115	58	0	58	440	9	449
	False	16	19	35	8	8	16	0	0	0	24	27	51
	Subtotal	286	25	311	120	11	131	58	0	58	464	36	500
Inorganic	True	273	7	280	111	4	115	50	0	50	434	11	445
	False	10	29	39	4	12	16	0	0	0	14	41	55
	Subtotal	283	36	319	115	16	131	50	0	50	448	52	500
Electric	True	188	11	199	76	9	85	61	0	61	325	20	345
	False	11	96	107	1	47	48	0	0	0	12	143	155
	Subtotal	199	107	306	77	56	133	61	0	61	337	163	500
Biochemistry	True	141	10	151	56	10	66	70	0	70	267	20	287
	False	14	137	151	7	55	62	0	0	0	21	192	213
	Subtotal	155	147	302	63	65	128	70	0	70	288	212	500
Physiology	True	176	19	195	80	4	84	93	0	93	349	23	372
	False	20	66	86	7	35	42	0	0	0	27	101	128
	Subtotal	196	85	281	87	39	126	93	0	93	376	124	500
Gastroenterology	True	215	5	220	95	3	98	29	0	29	339	8	347
	False	16	91	107	9	37	46	0	0	0	25	128	153
	Subtotal	231	96	327	104	40	144	29	0	29	364	136	500
All fields	True	1263	58	1321	530	33	563	361	0	361	2154	91	2245
	False	87	438	525	36	194	230	0	0	0	123	632	755
	Subtotal	1350	496	1846	566	227	793	361	0	361	2277	723	3000

Table 6. The performances of true/false prediction for the sample articles.

Field		Recall	False recall	Precision	Accuracy
Condensed matter	Training set	0.978	0.543	0.944	0.929
	Testing set	0.974	0.500	0.933	0.916
	Cited=1 or Cocit=1	1.000	-	1.000	1.000
	Whole sample	0.980	0.529	0.948	0.934
Inorganic	Training set	0.975	0.744	0.965	0.947
	Testing set	0.965	0.750	0.965	0.939
	Cited=1 or Cocit=1	1.000	-	1.000	1.000
	Whole sample	0.975	0.745	0.969	0.950
Electric	Training set	0.945	0.897	0.945	0.928
	Testing set	0.894	0.979	0.987	0.925
	Cited=1 or Cocit=1	1.000	-	1.000	1.000
	Whole sample	0.942	0.923	0.964	0.936
Biochemistry	Training set	0.934	0.907	0.910	0.921
	Testing set	0.848	0.887	0.889	0.867
	Cited=1 or Cocit=1	1.000	-	1.000	1.000
	Whole sample	0.930	0.901	0.927	0.918
Physiology	Training set	0.903	0.767	0.898	0.861
	Testing set	0.952	0.833	0.920	0.913
	Cited=1 or Cocit=1	1.000	-	1.000	1.000
	Whole sample	0.938	0.789	0.928	0.900
Gastroenterology	Training set	0.977	0.850	0.931	0.936
	Testing set	0.969	0.804	0.913	0.917
	Cited=1 or Cocit=1	1.000	-	1.000	1.000
	Whole sample	0.977	0.837	0.931	0.934
All fields	Training set	0.956	0.825	0.936	0.921
	Testing set	0.941	0.843	0.936	0.913
	Cited=1 or Cocit=1	1.000	-	1.000	1.000
	Whole sample	0.959	0.837	0.946	0.929

Of the true articles, 95% were identified correctly, while only more than 80% of the false articles were correctly identified as false. The precision is about 95% and the accuracy is 90-95%. There are no extraordinary variations among the fields, except a low false recall ratio for the field

"Condensed matter". The performances for the testing set are comparable to those for the training set. Significant difference is not observed between the performances for the articles with *AutMatch* = 0 and those for *AutMatch* = 1. The performances are considerably improved by using information of citation of the sample articles by the source articles (*Cited*) and cocitation between the sample and source articles (*Cocit*). Totally, our starting objective, correct answer ratio of at least 90% was attained.

It should be noted, however, that the performances demonstrated here are based on only the retrieved articles that passed the first filtering stage. Since this stage, as previously explained, addresses its main aim to reduce huge amount of retrieved articles (including false articles with a high rate) to a manageable article set, a considerable number of "false-like" true articles might be eliminated at this stage. We cannot say what degree of performance would be attained if our logistic regression model were applied to the overall data set of retrieved articles, but can say the performances we obtained apply to the article set including "true-like" articles with relatively high rate.

(5) Determination of the regression model for discriminating all retrieved articles

Since it was shown that the regression model obtained from the training set worked well for the testing set as well as the training one, a final regression model based on Equation (1) for each of the six fields was determined using all the sample articles in the field (excluding those with *Cited* = 1 or *Cocit* = 1). The performances obtained were almost same as those for the training sets; the overall recall ratio, precision and accuracy across the fields were 95%, 93% and 92%, respectively.

Concerning the high performances obtained here, one might raise the question that, in general, performances would apparently look good in spite of the model used since even a few rough rules would apply well to the easy cases and the majority of cases are relatively easy. So, we compared the performances for 'easy' cases and for 'difficult' cases by the following way. As stated in the Subsection "Second Filtering Stage" of the Section "Method for Discriminating True and False Articles", each sample article was decided as true or false by two judges and when the decisions by the two did not agreed another judge made the final decision. We regarded as 'easy' when the two judgments agreed and as 'difficult' when disagreed. Thus, 'easy' cases were those in which both the two judgments were true or both were false depending on the final judgment as true or false, and 'difficult cases were those in which one judgment was true and another was false. The correct answer ratios of discrimination based on our final logistic regression models were calculated for 'easy' and 'difficult' cases. The results are shown in Table 7. In the set of sample articles finally judged as true, 97% of the 'easy' articles were predicted correctly as true while the correct answer ratio for the 'difficult' articles was 88%. In the set of sample articles finally judged as false, the correct answer ratios were 85% and 70% for the 'easy' and 'difficult' cases, respectively. Although the correct answer ratio for the ‘difficult’ cases is certainly lower than for the ‘easy’ cases, the difference is not drastic.

Table 7. Correct answer ratios for 'easy' and 'difficult' cases.

Manually judged as true				
Case	Judgements by two judges	#Articles	#Articles correctly predicted	Correct answer ratio
Easy	Both true	1998	1939	0.970
Difficult	One true/one false	247	218	0.883
Total		2245	2157	0.961
Manually judged as false				
Case	Judgements by two judges	#Articles	#Articles correctly predicted	Correct answer ratio
Easy	Both false	651	556	0.854
Difficult	One true/one false	104	73	0.702
Total		755	629	0.833

The Final True/False Prediction in the Second Filtering Process

Of 106,163 retrieved articles having passed the first filtering stage, 12,947 articles (12.2%) were cited by, or cocited with, its source article (*Cited* = 1 or *Cocit* = 1) and therefore discriminated as true. For remaining 93,216 articles, true/false prediction was carried out using the regression models for the individual fields, as defined in (5) of the previous subsection. Articles were considered true if the calculated *p* was less than 0.5 and false otherwise.

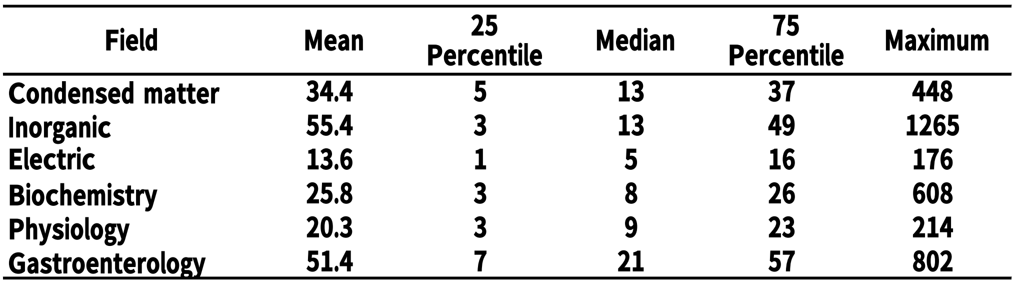
The results of the final true/false prediction at the second filtering stage are shown in Table 8. The number of true articles was found to be 90,052, which corresponds to 85% of the total number of retrieved articles having passed first filtering.

Table 8. The results of true/false discrimination at the first and second filtering stages.

Field	#Source authors	Passed first filtering			Passed second filtering			Passing rate (%)		
		AutMatch = 0	AutMatch = 1	Total	AutMatch = 0	AutMatch = 1	Total	AutMatch = 0	AutMatch = 1	Total
Condensed matter	357	8686	4853	13539	7428	4853	12281	85.5	100.0	90.7
Inorganic	422	15206	9327	24533	14072	9327	23399	92.5	100.0	95.4
Electric	379	7420	1824	9244	3641	1526	5167	49.1	83.7	55.9
Biochemistry	476	12203	5583	17786	7031	5254	12285	57.6	94.1	69.1
Physiology	400	5380	3886	9266	4249	3853	8102	79.0	99.2	87.4
Gastroenterology	561	18029	13766	31795	15052	13766	28818	83.5	100.0	90.6
All fields	2595	66924	39239	106163	51473	38579	90052	76.9	98.3	84.8

The predicted average, quartiles and maximum numbers of true articles by the source authors are shown in Table 9. The averages and quartiles can be considered reasonable since our search period (30 years) sufficiently covers a typical researcher’s productive life time. The maximum value, however, appears somewhat too large, suggesting that for some source authors, articles by homonymous authors have not been sufficiently eliminated.

Table 9. The mean, quartiles and maximum values of articles per source author.



Field	Mean	25 Percentile	Median	75 Percentile	Maximum
Condensed matter	34.4	5	13	37	448
Inorganic	55.4	3	13	49	1265
Electric	13.6	1	5	16	176
Biochemistry	25.8	3	8	26	608
Physiology	20.3	3	9	23	214
Gastroenterology	51.4	7	21	57	802

Conclusions

An initial search on 2,595 source authors yielded 629,000 retrieved articles. Of these, 106,000 (17%) passed first filtering and 90,000 (14%) were identified as true articles through second filtering. A large number of articles were eliminated by first filtering, but most of the remaining articles were not eliminated by second filtering. The algorithm used for second filtering was investigated carefully through manual judgment of the sample articles, but the criteria for first filtering relied only on rough inspection. Given these facts, since first filtering focused on eliminating false articles and reducing a huge number of articles, a considerable number of true articles might be eliminated. In other words, there might be a certain number of true articles eliminated because of completely different addresses from the source article or little citation relationship between journals of the source and retrieved articles.

The results of the discrimination of sample articles showed that our original aim is achieved, that is, more than 90% of the articles were correctly identified using the discrimination function based on a logistic regression model. But it has to be added that the performances are obtained for only the retrieved articles that passed the first filtering stage.

Existence of coauthor(s) of same name and address similarity were found to be the most important discriminating features, and title words similarity and the strength of citation relationships between journals were also highly effective for discrimination. Moreover, due to the extremely high incidence of homonyms in specific countries (China, Japan, South Korea, and Taiwan), it was important to consider whether the author was from one of these countries.

If a retrieved article was cited by, or cocited with, its source article, then it was found to be almost certainly true. Therefore, that information is particularly effective for prediction. Bibliographic coupling between a retrieved article and its source article would also strongly imply the retrieved article is true although it was not used in this study because cross-checking of a huge amount of reference data had to be done.

The method proposed in this paper used as various features available from the WoS database as possible. They included not only features based on bibliographic data such as *AutMatch*, *Add_Sim*, *Tit_Sim*, *Age* and *FEA* but also those based on citation data such as log (X+1), *Cited* and *Cocit*. In contrast to this, many of the author disambiguation approaches previously reported, including those by Giles group (Han et al. 2004; Han, Zha & Giles 2005; Huang, Ertekin & Giles 2006; Song et al. 2007), McCallum group (McCallum & Wellner 2003; Kanani, McCallum & Pal 2007; Kanani & McCallum 2007) and Cota et al. (2010), addressed their aim mainly to develop a more advanced clustering algorithm rather than to seek effective features for disambiguation.

On the other hand, the methodology by Torvik's group (Torvik et al. 2005; Torvik & Smalheiser 2008) pursued both the aims. Of the features they used for discriminating same-named authors in the MEDLINE database, they found that common coauthor information was the most important, journal name and author's second name initial were the next, and followed by affiliation words, title words and MeSH terms (Torvik et al. 2005). (In addition, they showed author's e-mail address and full first name were very effective if they were available (Torvik & Smalheiser 2008).) Those results are similar to our results to some extent although we did not use second name initial and MeSH terms and we use citation relationship between journals instead of journal name Torvik et al. used. We utilized the citation or cocitation relationship which was not used by Torvik et al., showing it was valuable information for author discrimination. The approach proposed by Torvik's group exploited rigorous clustering method considering the problem of transitivity violations into account as well as the rich discriminating features, leading to very high performances. On the other hand, we did not adopt a clustering method which calculates similarities of all article pairs in a given data set but an approach comparing individual retrieved articles with their source articles. Although this approach has some limitations such as an inability of solving the transitivity problem, it can process a large number of data with fewer steps. Therefore, we think the methodology proposed in this study would be suitable for a situation we faced, in which 'true' articles by target authors have to be identified, with certain accuracy (not in perfect), among a very large number of articles contaminated by 'false' ones by many homonymous authors.

In this study, a common methodology (logistic regression analysis with the same independent variables) was applied to six different subject fields and many source authors of various affiliations, and the results obtained (the regression coefficients and their significance level and also the discrimination performances) did not greatly vary among the fields. This suggests the methodology is generally applicable beyond fields and affiliation variations. It may be a limitation of this study not to apply the method to data obtained from databases other than WoS. However, data concerning coauthors, addresses and countries of authors, title words and publication years are also available from other bibliographic databases, hence the results related to these features would be generalized. On the other hand, citation data used in the discriminating variables log (X+1), *Cited* and *Cocit* can be obtained only from the WoS and a few other databases with citation index. It is one of the distinctive points of this study to use such features since few studies are found demonstrating empirically that those features are effective for author disambiguation, except for one by Tang & Walsh (2010) which used bibliographic coupling information.

Acknowledgement

We thank Dr. Shigeaki Yamazaki, Professor of Aichi Shukutoku University and Dr. Keita Tsuji, Associate Professor of the University of Tsukuba, for their useful advices on this study. This work was supported by JSPS KAKENHI (18300080).

Notes

[1] The objective of that study is to identify the major factors that affect the number of citations received by a given article; one of these factors is supposed to be the number of articles already published by the author(s) of the article. To obtain such data, author name searches were carried out for a set of source articles.

References

Aksnes, D.W. (2008). When different persons have an identical author name. How frequent are homonyms? Journal of the American Society for Information Science and Technology, 59(5), 838–841.

Bornmann, L., & Daniel, H.-D. (2007). Multiple publication on a single research study: Does it pay? The influence of number of research articles on total citation counts in biomedicine. Journal of the American Society for Information Science and Technology, 58(8), 1100–1107.

Cota,R.G., Ferreira,A.A., Nascimento, C., Gonçalves,M., & Laender,A.H.F. (2010). An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. Journal of the American Society for Information Science and Technology, 61(9), 1853–1870.

Han, H., Giles, C.L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (pp. 296–305). NewYork: ACM Press.

Han, H., Zha, H., & Giles, C.L. (2005). Name disambiguation in author citations using a K-way spectral clustering method. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (pp. 334–343). NewYork: ACM Press.

Huang, J., Ertekin, S., & Giles, C.L. (2006). Efficient name disambiguation for large-scale databases. In Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases. Lecture Notes in Artificial Intelligence, 4213, 536–544.

Kanani, P., & McCallum, A. (2007). Efficient strategies for improving partitioning-based author coreference by incorporating Web pages as graph nodes. In Proceedings of theAAAI Sixth International Workshop on Information Integration on the Web (pp. 38–43). Menlo Park, CA: AAAI Press.

Kanani, P., McCallum, A., & Pal, C. (2007). Improving author coreference by resource-bounded information gathering from the Web. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (pp. 429–434). Menlo Park, CA: AAAI Press.

Kang, I.-S., Na, S.-H., Lee, S., Jung, H., Kim, P., & Sung,W.-K. et al. (2009). On co-authorship for author disambiguation. Information Processing & Management, 45(1), 84–97.

McCallum, A., & Wellner, B. (2003, August). Object consolidation by graph partitioning with a conditionally-trained distance metric. Paper presented at the KDD Workshop on Data Cleaning, Record Linkage and Object Consolidation, Washington, DC.

Moed, H.F. (2005). Citation analysis in research evaluation. Dordrecht, the Netherlands: Springer.

Rinia, E.J., van Leeuwen, Th.N., van Vuren, H.G., & van Raan, A.F.J. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria. Evaluation of condensed matter physics in the Netherlands. Research Policy, 27(1), 95–107.

Smalheiser, N.R., & Torvik, V.I. (2009). Author name disambiguation. Annual Review of Information Science and Technology, 43, 287–313.

Song, Y., Huang, J., Councill, I.G., Li, J., & Giles, C.L. (2007). Efficient topic-based unsupervised name disambiguation. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (pp. 342–351). NewYork: ACM Press.

Tang, L., & Walsh, J.P. (2010). Bibliometric fingerprints: Name disambiguation based on approximate structure equivalence of cognitive maps. Scientometrics, 84(3), 763–784.

Torvik, V.I., & Smalheiser, N.R. (2009). Author name disambiguation in MEDLINE. ACM Transactions on Knowledge Discovery from Data, 3(3), 11.

Torvik, V.I., Weeber, M., Swanson, D.R., & Smalheiser, N.R. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation. Journal of the American Society for Information Science and Technology, 56(2), 140–158.

van Raan, A.F.J. (2006). Performance-related differences of bibliometrics statistical properties of research groups: Cumulative advantage and hierarchically layered networks. Journal of the American Society for Information Science and Technology, 57(14), 1919–1935.

Wooding, S., Wilcox-Jay, K., Lewison, G., & Grant, J. (2006). Coauthor inclusion: A novel recursive algorithmic method for dealing with homonyms in bibliometric analysis. Scientometrics, 66(1), 11–21.