Scientometrics (2014) 99:823–838
DOI 10.1007/s11192-013-1214-2

Institution name disambiguation for research assessment

Shuiqing Huang · Bo Yang · Sulan Yan · Ronald Rousseau

Received: 2 August 2013 / Published online: 28 December 2013

Abstract Research evaluation is a necessity for management of academic units (scientists, research groups, departments, institutes, universities) and for government decision making in science and technology. Yet, wrong conclusions may be drawn due to errors in assignments of authors to institutions. To improve existing techniques of institution name disambiguation (IND) based on word similarity or editing distance, a rule-based algorithm is proposed in this study. One-to-many relationships between an institution and many variant names under which it is referred to in bylines of publications are recognized with the aid of statistical methods and specific rules. The performance of the rule based IND algorithm is evaluated on large datasets in four fields. These experimental results demonstrate that the precision of the algorithm is high. Yet, recall should be improved.

Keywords Institution name disambiguation (IND) · Rule-based system · Artificial intelligence · Informetrics

Introduction

Research evaluation is a necessity for management of academic units (research groups, departments, institutes, universities) and for government decision making in science and

S. Huang · B. Yang (✉) · S. Yan
College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, Jiangsu, People's Republic of China
e-mail: mail.boyang@gmail.com

R. Rousseau
Institute for Education and Information Sciences, IBW, University of Antwerp (UA), Venusstraat 35, 2000 Antwerp, Belgium

R. Rousseau
KU Leuven, 3000 Louvain, Belgium

technology. Also on a personal level researchers are evaluated in relation to tenure and promotion decisions. Although research evaluation consists of much more than the evaluation of the number and quality of published articles in this contribution we will only pay attention to this particular aspect. Moreover we focus on scientometric methods as performed by outsiders. By this we mean that we act as if the evaluator does not have access to a detailed and exhaustive list of all publications of all scientists and all research units under study. A typical case in which this approach is taken is when the staff of Thomson Reuters develops the Essential Science Indicators (ESI), in which the top 1 % institutions, authors, countries and journals are ranked within each subject by citation data taken from Thomson Reuters' databases. As use of these lists not only in theoretical and informetric studies, but also in the practice of scientific evaluation is not uncommon (Csajbók et al. 2007) a critical scientific approach consists in checking (at least partially) if data, and hence the rankings, as provided in the ESI are correct. However, a large-scale analysis that we performed shows that the lists provided in the ESI are not as reliable and accurate as one might expect. It seems that author names and institution names are not always in one-to-one relation with persons and institutions. Even on country level ambiguities exists (for example: what is the relation between Wales, Northern Ireland and the United Kingdom, or between Hong Kong, Macau and the P.R. China?) Such ambiguities may lead to deviations in data and discussions in the evaluation of scientific outputs at the country level (Narin et al. 1988). Taking agricultural science as an example, the data used for our study shows that nearly 25 % of the author names (consisting of a surname and initials) are shared by at least two different individuals. When authors or publishers do not provide unambiguous data then it becomes very difficult for data aggregators such as Web of Science (WoS) or Scopus to provide accurate data or at least statistically reliable ones. Yet this is what we try to achieve. The data from the ESI, WoS and Scopus has been widely applied in many studies, though few refer to the correctness of the data and the resulting reliability (or lack thereof) of the conclusions based on possibly incorrect data. In a study by Taşkın and Al (2013) it was shown that wrong affiliation information happens often for hospitals (concretely in Turkey but a case study on English university hospitals indicates that this specific problem is not only a Turkish phenomenon (Praal et al. 2013)) and how this leads to misrepresentations in collaboration maps. Moreover, the number of highly-cited authors plays an important role in determining the position of an institution in a university ranking list, so incorrect author names (e.g. one author represented as two persons so that none of the two enters the top 1 %) as well as wrongly spelled or unnormalized institution names may influence such university rankings (Van Raan 2005).

All studies based on publications or citations need perfect data, or at least as perfect as humanly possible. Whether the aim is the evaluation of scientists, research groups or institutions, or the construction of networks, author or institution name disambiguation (AND—IND) is indispensable (Yang et al. 2008; Strotmann and Zhao 2012). Although the title of an article is one of the important indicators to recognize an author's identity, some scientists are involved in several research topics so that keyword based clustering techniques for author disambiguation may not work. In this kind of situations, affiliation information can be helpful to identify authors who publish papers on different topics. A reliable table mapping various institution names to a single entity is the first priority to fulfill the task of institution evaluation (De Bruin and Moed 1990). During the production process institution information in the WoS is often abbreviated but not (fully) normalized. Abbreviations are usually based on the form provided by authors in their publications. For various reasons authors from the same institution often provide different spellings of the institution's name and sometimes it even happens that the same author uses different

versions of his/her institution name in different articles. This phenomenon is common in non-English countries as institution names are given in the local language and can be translated into English in different ways. In addition, sometimes an address (in the WoS) begins with the university name, followed by a faculty name or the name of a school; sometimes these metadata are given in another order, or given only partially. All this may lead to incorrect research results.

A study of the metadata provided by the WoS shows that often basic structures, e.g. university—department—street address, city, including ZIP code, can be recognized. To improve IND an algorithm for institution name mapping is proposed in this study. Some many-to-one relationships are recognized with the aid of statistical methods and some extra rules to achieve the aim of IND are applied. Our approach can be described as a rule based statistical method. Finally, the performance of the algorithm is evaluated using recall and precision values.

Related work

From the perspective of scientific evaluation, publication metadata contain various entities that must be known in an unambiguous way, such as country names, author names and institution names. Because of different writing styles and coding rules (Richardson 2010), some ambiguity seems to be unavoidable. Especially, the correct identification of author and institution names is complicated. Moreover, these two types of entities are often related. The scientific literature contains plenty of information that can be used for AND, such as title of publications, abstracts, keywords, journal names, coauthor names, street addresses and email addresses (D'Angelo et al. 2011; Levin et al. 2012), while the amount of direct evidence for IND is more limited. IND can be used as a part of AND, hence institution name mapping is an indispensable step for many AND projects.

The essence of IND is to map an institution name to its unique name. One can discern two types of methods to reach this aim: (1) using a look-up table drawn by an official office which contains all scientists in the country with full details as to the university or universities to which each scientist belongs; (2) institution name clustering by word similarity. Abramo et al. (2011) study of the Italian university system used a staff-university table maintained by CINECA (http://www.cineca.it/en/content/about-us) on behalf of the Ministry of Education, Universities and Research including information for each scientist in an Italian university, such as discipline, department and position. It is an example of the first type of approach to assign articles to institutions, namely via its authors. As keywords extracted from institution names may indicate characteristics of an institution Morillo et al. (2013) made use of this information to classify a group of Spanish institutions extracted from Spanish scientific publications into nine predefined sectors. Similarly in Onodera et al. (2011) the frequencies of all words appearing in institution names in the given dataset were calculated and each word was correspondingly weighted. Then the total weight of the shared words between two institution names was used to measure the similarity of the two names. Another clustering algorithm called normalized compression distance (NCD) was tested by Jiang et al. (2011). Their experiments showed that to cluster different names referring to the same institution into one group it is an effective technique. As a general technique for AND, edit distance can be used to measure the similarity of two strings by the number of operations required to transform one string into the other (Torvik et al. 2005; Smalheiser and Torvik 2009). The edit distance, also called Levenshtein distance, between two strings is defined as the minimum number of edits needed to transform one string into

the other, where edit operations are insertion, deletion, or substitution of a single character (Levenshtein 1966). A string clustering strategy proposed by French et al. (2000) based on the concept of edit distance was tested with a collection of institution names and the result confirms the effectiveness of the technique on institution address clustering. In some cases there are different address formats for the same institution, so a unification method based on parameterized finite-state graphs was proposed by Galvez and Moya-Anegón (2006, 2007). Although it has nothing to do with IND, this study makes it possible to extract automatically the name of the chief institution (say, the university) from institutional addresses.

Yet, IND is not always applied necessary for AND (although we think that AND always benefits from IND). Indeed, web mining techniques and information extraction can be applied to AND to recognize an author with the aid of clues from the Web other than publications (Pereira et al. 2011; Bollegala et al. 2012; Tang et al. 2012). However, such techniques may not always work for IND.

If a scientist-institution list manually maintained in a unified format by an official scientific administration exists this is the best approach: it is formal and can be accurate. Yet, building and updating such a list is a difficult task. Moreover, different writing styles, by authors or required by publishers, make that even this approach is not straightforward. Hence, in practice this approach is of limited application. Although the word similarity based method has been proved to be an effective technique to cluster institution names, it is clear that also this approach has no universal applicability. In some cases two strings have a high word similarity or the edit distance is short, yet, the word strings do not refer to the same institution, e.g. Zhejiang University of Technology and Zhejiang University of Science and Technology, are two different universities, both located in Hangzhou, Zhejiang province. By contrast, some pairs of strings with a low value of word similarity or a long edit distance may refer to the same institution. This may happen when a university has changed its name e.g. at the occasion of a merger. We conclude that measuring the affiliating relationship between two institution names simply by the values of word similarity or edit distance is not fully reliable.

Research questions

Most AND investigations fall into one of the following situations: a single author publishes articles under different names: the case of synonyms (in our field the many variants of Derek J. de Solla Price's name spring to mind) or several individuals share the same name: the case of homographs (Egghe and Rousseau 1990, p. 218; Smalheiser and Torvik 2009; Cota et al. 2010). These two situations also occur in IND investigations. For reasons of different writing styles, or institutional transitions such as merging, a single institution sometimes has several names such as "Huazhong Normal University" and "Central China Normal University" (*Hua zhong* means Central China). Such cases are referred to as "single to multiple" (a single institute is referred to under multiple names) in short STM. On the contrary, occasionally some different institutions from regions with a similar language, culture and history share the same name (the Multiple to Single case). For example, there are two different universities called "Soochow University" one in Mainland China and one in Taiwan. Such problems may also occur during translation: in Belgium Katholieke Universiteit Leuven and Université catholique de Louvain (situated in different cities) can both be translated in English as Catholic University of Louvain. As the former situation is more common, this study focusses on how to discover and solve

efficiently the STM phenomenon, including those that occur by simple writing errors, from a large collection of scientific literature so as to improve the reliability of literature-based institution evaluation.

Solving the STM problem

By the term 'solving the STM problem' we mean partitioning institutional names as they occur in a subset of articles in a database (in this contribution this database is Thomson Reuters' WoS) in non-overlapping groups such that each group refers to a unique institution in the real world and assigning a unique name to each group. One such group resulting as a solution of the STM problem will be called an institution synonym set, ISS in short.

The main research questions of this contribution are:

1. To construct a rule-based algorithm to solve the STM problem.
2. To show to which extent the rule-based algorithm proposed in this study is capable of solving the STM problem.
3. To investigate the reliability of this technique in different fields. In other words: how accurate are the resulting ISSs?

Method

Data production such as performed by Thomson Reuters divides an article's address into several sections such as: university, department, street, postcode, region and country. Although there exist multiple cases of the STM phenomenon, one may distinguish some patterns that often occur. Generally a pair of institution names possibly refers to the same institution, if the word similarity between them is high. However, word similarity based indicators are not reliable enough for IND because of the occurrence of similar naming patterns for different research institutions, especially for universities. Hence we want to detect rules to detect patterns in which authors write their own name and that of the institution(s) they belong to, in order to construct a rule-based technique for solving the STM problem. Our approach is introduced in detail in this section.

Classification of STMs

To guarantee the reliability of IND techniques and to link different names with their real (unique form) institution, a thorough analysis on the causes of the STM phenomenon is essential. We distinguish five main reasons (see also Table 1):

(1) Translation. It happens in non-English speaking countries that several translated names are used synchronously or alternatively. For example, the term "liaoning univ petr & chem technol" has been used by some authors even though "liaoning shihua univ" is the official name of Liaoning Shihua University (shiyou, abbreviated to shi, means petroleum, and huaxue, abbreviated to hua, means chemistry).
(2) Spelling. Some Asian languages such as Chinese and Japanese use symbols (characters) that do not represent one sound (as most letters in Western languages) but one word or concept. Some words may consist of two or more characters and when transliterating one may or may not write these two characters as one word. In

Table 1 Forms of the STM phenomenon and examples

Types	Samples
Translation	"kangmung natl univ" and "gangneung wonju natl univ" (Korea); "liaoning shihua univ" and "liaoning univ petr & chem technol" (China);
Spelling	"shanghai jiaotong univ" and "shanghai jiao tong univ" (China)
Institution transition	"kharkov am gorkii state univ" is an old name of "kharkov natl univ" (Ukraine); "jilin univ" and "jilin univ technol" (China)
Spelling errors	"yuan univ" should be "yuan ze univ" (Taiwan)
Institution and divisions	"univ Strasbourg" and its division "univ louis pasteur Strasbourg 1"

Chinese pinyin words are usually written as a whole. For example: one writes Shanghai (two characters), daxue = university (two characters), kexuejia = scientist (three characters). However this is not always clear cut. An example is the case of Shanghai Jiao Tong University. As the word jiaotong, meaning communication or transportation, is an existing word one often encounters the form Shanghai Jiaotong University instead of the official form, see http://en.sjtu.edu.cn/about-sjtu/overview. Moreover, diacritics, as used in many western languages, but also in Arabic, Cyrillic and Chinese pinyin (to mark tones), are always removed in the WoS, enlarging the scope for confusion.

(3) Institutional transition. Many universities and research institutions have a long history, during which they changed names. Sometimes these name changes are the result of a merger, sometimes it is just a way to show a change in strategy. Yet it often happens that several years after the name change researchers still use the old name. In Table 1 we show the case of Jilin University (Jilin University of Technology merged with Jilin University in 2000).

(4) Spelling errors. Spelling errors are quite common in the address field of the WoS, especially in the main parts of institution names. This kind of error is partly introduced by the authors themselves, partly during database production. For instance, the Chinese city Xi'an, capital of Shaanxi Province in China, is usually indexed as Xian in the WoS.

(5) Institution versus divisions. Some authors provide the university name as the chief institution name in their articles while some others only list departments, research divisions or campus names in their articles. Although the recognition of the exact affiliation relation between an institution and its divisions is not really an STM phenomenon, in order to have reliable statistics it is necessary to find the institute(s) of articles published only under the name of a division. Hence in this section this phenomenon too is considered a form of the STM phenomenon.

Heuristics for rule-based institution name disambiguation

When studying the publication of one author one may (rightly) think that if two institution names are similar to some extent and their division names or postcodes are identical, then probably the two institution names refer to one institution. Consider for example the author "Diao, KF". For this author one finds two different institution names, namely:

1. Linyi Normal Univ, Dept Math, Linyi 276005, Shandong, Peoples R China.
2. Linyi Univ, Sch Sci, Linyi 276005, Shandong, Peoples R China.

The institution names in these two addresses are "Linyi Normal Univ" and "Linyi Univ", while "Dept Math" and "Sch Sci" are division names. Besides country and province, also postcodes are identical ("Linyi 276005"), so it is very likely that the two names are a case of the STM phenomenon (in short: STM candidates) and stand for the same institution. Actually Linyi Normal University has been renamed Linyi University since 2010. In other cases, when postcodes are missing, some STM candidates can be identified because they have identical division names. This is one example of a heuristic we will apply.

Procedures of rule-based institution name mapping

We recall that the purpose of IND is to partition institutional names as they occur in a subset of articles from the WoS in non-overlapping groups, i.e. ISSs, such that each group refers to a unique institution in the real world and assigning a unique name to each ISS. The whole process is implemented in three steps.

Step 1: Building an author-institution table
Step 2: Recognizing STM candidates based on author blocks
Step 3: Frequency based institution name mapping

Step 1: Building an author-institution table

Author names can be useful resources to check if several institution names are potential STM candidates. The first step in our procedure for IND is to extract author names and their institution names and build an author-institution table. In this table the author name is abbreviated as surname and initials as provided in the WoS, e.g., "Li, Y" or "Jin, BH". Perfect duplicates are removed and if an author has two or more institutions they are all mentioned.

For example, one obtains, based on one publication:

Glanzel, W
Katholieke Univ Leuven, Ctr R&D Monitoring ECOOM, Dept MSI, B-3000 Louvain, Belgium
Hungarian Acad Sci, IRPS, Budapest, Hungary

In the author-institution table, we include:

Authors	Institutions
Glanzel W	Katholieke Univ Leuven :: Hungarian Acad Sci

Moreover, based on another publication, one obtains:

Glanzel W.
Katholieke Univ Leuven, Ctr R&D Monitoring ECOOM, B-3000 Louvain, Belgium
Katholieke Univ Leuven, Dept MSI, B-3000 Louvain, Belgium

Hungarian Acad Sci, IRPS, Budapest, Hungary

Whether or not this is the same Glanzel, W. this does not change the author-institution table, as, in this step, we do not take departments or any other institutional division into account. In this particular case these data refer to the same Glanzel W but in the case of Chinese names a name such as Yang B. (just an example) can refer to different scientists. Of course similar cases of homonyms also occur often in Korea (Kim and Cho 2013), Japan, India and in most other countries.

An example of this case is the following:

Yang, B. (full name: Yang Bo, but not an author of this article)
Nanjing Univ, Inst Clin Lab Med, Nanjing Jinling Hosp, Clin Sch, Med Coll, Nanjing 210002, Jiangsu, Peoples R China

This leads to the following entry in the author-institution table:

Authors	Institutions
Yang B.	Nanjing Univ

In another article we find:

Yang B. (full name: Yang Bin)
Nanjing Univ, Dept Urol, Sch Med, Affiliated Drum Tower Hosp, Nanjing 210008, Jiangsu, Peoples R China

Although this is another person, it does not change the author-institution table. The reason is that we are performing IND, not AND. However, when we encounter:

Yang B. (full name: Yang Bo)
Nanjing Agr Univ, Nanjing, Peoples R China

this leads to a new entry:

Authors	Institutions
Yang B.	Nanjing Agr Univ

Although in this first step we only use author-institution tables, we will (of course!) also make use of other address information. An address is subdivided into segments, distinguished by commas. In this way we obtain, for example the address: "Linyi Normal Univ (segment 1), Dept Math (segment 2), Linyi 276005 (segment 3), Shandong (segment 4), People's R China (segment 5)". The last segment is called the country, although it may actually be a region or a combination of a state and a country. We even found the following last segment consisting of a state, a ZIP code and a country, such as in:

SUNY Upstate Med Univ, Dept Neurosci & Physiol, Syracuse, NY 13210 USA

Step 2: Recognizing STM candidates based on author blocks

For the reasons summarized in Table 1 many forms of the STM phenomenon occur in the literature, so recognizing a set of STM candidates is the most important procedure for IND

research. In this study, an author block based algorithm is applied to Same Institution Recognition (Levin et al. 2012). The table of author-institution is separated into blocks by author names, where each block consists of an author name, e.g. Glanzel W. or Yang B. and all institution names associated with this author name. STM recognition is then performed within an author block. Although it may not be true that all institution names within a block refer to the same institution, certainly some STM candidates are found in this way. In a block we recognize three situations: (1) Different individuals share an author name and belong to different institutions; (2) One individual publishes articles with different institution names that may refer to the same institution and are just variations of the same institution name; (3) Different individuals share the same author name and belong to the same institution. As explained above, this third alternative does not matter in the case of IND. A set of STM candidates composed of institution names identified in the second situation is vital for rule-based IND. Therefore, first some pairs of institution names in a block will be filtered out by word similarity of name strings and then some of the remaining pairs are recognized as STM candidates based on structural patterns in the addresses (see further). Concretely we start from the set AA consisting of all (unordered) couples of addresses occurring in a given author block. In a first phase some couples will be chosen and placed in a subset C. In a second phase this subset will be reduced to another subset called D (see Fig. 1).

Rules

We recall that we always work within one author block. For a given author block, we apply the following rules for same institution recognition. The symbols N1 and N2 denote the words or word fragments making up the first segment of an institutional address, i.e. the part shown in an author-institution address; for example: N1 = Nanjing Univ, N2 = Nanjing Agr Univ. We systematically consider all (N1;N2) pairs. Hence, if there are 5 addresses in an author block then we consider 10 cases.

Rule1: If N1 and N2 have the same number of words then the following rule is applied. Denoting the number of shared words, i.e. {N1} ∩ {N2}, by shared(N1;N2) and the number of distinct words in {N1} U {N2} by total(N1;N2) then the similarity between N1 and N2 is denoted by Sim(N1;N2) and obtained as:

$$\mathrm{Sim}(N1;N2) = \mathrm{shared}(N1;N2) \,/\, \mathrm{total}(N1;N2)$$

This is nothing but the well-known Jaccard measure (Jaccard 1901). It is a relative similarity measure. If the similarity is larger than 0.6 then the pair (N1;N2) is placed in C, a subset of AA. An example: if N1 = Katholieke Univ Leuven and N2 = Hungarian Acad Sci then shared(N1;N2) = 0, total(N1;N2) = 6 and Sim(N1;N2) = 0; if N1 = Antwerp Univ and N2 = Univ Antwerp then shared(N1;N2) = 2, total(N1;N2) = 2 and

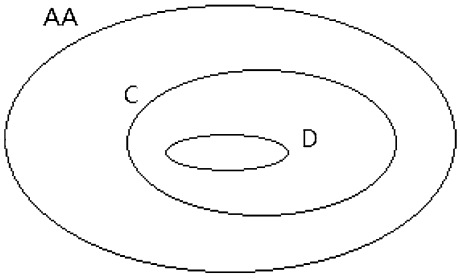

Fig. 1 Schematic representation of reduction of all possible pairs within an author block

Sim(N1;N2) = 1; if, however, N1 = Antwerp Univ and N2 = Univ Antwerpen then shared(N1;N2) = 1, total(N1;N2) = 3 and Sim(N1;N2) = 1/3.

In particular, if N1 and N2 contain the same words in another order then the Jaccard similarity is equal to one. Another example is: N1 = ilam univ and N2 = univ ilam.

Rule2: If the numbers of words of N1 and N2 are not equal and shared (N1;N2) $\geq$ 2, add the pair to C. In contrast to Rule1 this is not a relative but an absolute similarity rule. We will refer to it as the share rule (here with threshold 2). An example: if N1 = Nanjing Univ and N2 = Nanjing Agr Univ then shared (N1;N2) = 2 and this pair is added to the subset C.

Rule3: If one of N1 and N2 is a substring or abbreviation of the other, add the pair to C. Example are; N1 = Univ Antwerp and N2 = Univ Antwerpen (substring) and N1 = Univ Antwerp and N2 = UA (abbreviation).

Rule4: If the edit distance of N1 and N2 is less than 0.2, add the pair to C. An example where this rule can be applied is N1 = Islam Azad Univ and N2 = Islamic Azad Univ. The edit distance between strings is calculated by a modified version of the Jaro-Winkler algorithm (Alias-I 2002).

When these four rules are applied for all pairs, the next rule is applied, removing some pairs that certainly do not refer to the same institution.

Rule5: If in the current author block two addresses containing N1 and N2 in the set C share the same country name then keep N1 and N2 in C, otherwise remove the pair. Note that at this point we go beyond the information as shown in the previous author—institution tables and also use country (=last segment) information.

Rule6: After the filtering procedure of rule5, we continue with the pairs in the subset C, using the full address information if necessary. If the number of segments in an address containing N1 is not equal to that of an address containing N2 or they are the same but the number of segments is at most three, and the second segments following N1 and N2 in the two addresses are identical, then put the pair (N1;N2) in a subset D of C. If there are more than three segments and one of the segments (with the exception of the first, second or last segment) are exactly the same then add the pair (N1;N2) to D. A typical case occurs when the postal codes are identical.

Although there are international research institutions with an overseas division, generally similar institution names belonging to the same country could indicate an STM case. So rule5 aims to reduce the possibility of STM mismatching between two institutions that share the same name and come from different countries. Structured address information is used in rule6 to filter out some unqualified STM candidates in the set C. If an institution name is similar to another and their division names, street names or postcodes are exactly the same, they could be members of the same ISS. In case of short institution names consisting of the same number of segments (but not more than three) the division names (the second part of an address) will be compared directly in rule6.

The six rules proposed above can be divided into three modules, where rules from 1 to 4 fall into the first module, rule5 is the second and rule6 is the third. The three modules must be executed one by one, and finally a new set D is produced in this step.

As an author block may contain several ISSs the above procedure is applied several times. Concretely we propose three rounds of filtering.

Step 3: Frequency based institution name mapping

When step 2 is applied for each author block, we have a D for each author block. A given pair (N1;N2) may or may not appear in more than one D set. To improve the accuracy of same institute recognition, we may apply a threshold T. Concretely we may use only pairs

(N1;N2) that occur in at least two D sets (taking T = 2). Then clusters are formed: if (N1;N2) is a pair occurring in at least two D sets (if we take T = 2) and (N1;N3) is another one (also occurring in at least two D sets) then the cluster {N1,N2,N3} is formed. Each cluster is an ISS as obtained by our rule-based algorithm. One of the names will be chosen as the formal institution name and the other ones are its informal names. Usually official names are short so we chose the shortest one, preferring MIT above Massachusetts Inst Technol (if there is more than one shortest name, just select one randomly).

In this step, the thresholds can be adapted depending on the application one has in mind. A high threshold may catch the most common names for an institution while some unusual names (typically printing errors) can only be recognized when a low threshold (T = 1) is chosen.

This procedure leads to a result as shown in Table 2.

Experiment

In this section we describe an experiment preformed to verify the performance of our algorithm.

Data source

Although there is no direct relation between the data provided in the WoS and the ranked lists published in the ESI; and the categories used for the ESI are different from the Web of Science categories, the basic data sources used for the ESI and the WoS are the same. Hence the data set is built based on ESI categories to make the result of this study (namely, finding unique institution names) comparable with institutions as used in the ESI rankings. To evaluate the performance of the IND algorithm proposed in this study in different fields, the algorithm is applied to four fields namely mathematics, computer science, psychology and economics and business. First a journal list for each field as defined in the ESI is obtained, and then all metadata of articles published from 2008 to 2011 is exported from the WoS to build an initial data set. Articles published before 2008 are ignored because author names in these articles are not matched to an address. Data collection was done

Table 2 Five examples of ISSs

Formal name	Informal name	Country
Univ jj strossmayer	jj strossmayer univ osijek::univ jj strossmayer osijek::univ josip juraj strossmayer osijek::josip juraj strossmayer univ osijek::univ j j strossmayer osijek::jj strossmayer univ	Croatia
irta	ctr udl irta::irta food technol::udl irta::irta ctr tecnol alimentos::irta finca camps & armet::irta monells::udl irta	Spain
univ bordeaux 2	univ victor segalen bordeaux 2::univ victor segalen bordeaux ii::univ victor segalen bordeaux	France
univ antwerp	univ instelling antwerp::univ antwerpen ua::univ antwerpen cde::univ antwerp vib::univ antwerp uia	Belgium
sun yat sen univ	sun yat sen zhongshan univ::zhongshan sun yat sen univ	China

between August 28, 2012 and September 3, 2012, leading to 111,999 articles in mathematics, 148,261 in computer science, 168,702 in psychology and 109,247 in economics and business.

Results

According to the IND algorithm mentioned above, the STM candidates in the four fields are saved in a batch of log files after the first two steps of data processing and then the next step is to generate the final collection D. In the third step if the frequency threshold T is 1, this means that all candidates recognized in the previous two steps will be included in the final set. For the sake of reducing noisy data, the threshold is set to 2 in this experiment. It follows that only pairs appearing more than once will be included in a new collection named F2. In the end the four F2 sets contained 366 institutions (ISSs) in mathematics, 425 in computer science, 737 in psychology and 228 in economics and business.

Presumably because of similar cultural backgrounds, the situation that different individuals share the same name happens frequently in Asian countries with many publications, such as China, Korea and Japan. Moreover, in these countries, elements constituting an institution name are often similar too. Hence the percentage of mismatched STM candidates is sometimes rather high. For this reason having the same postcode is required for an STM candidate in those countries. This improves the accuracy when rule5 is applied. Furthermore, the accuracy may be affected when the names of university systems are mixed with university names (Van Raan 2005). For example, in France university names in a university system are similar to each other and mismatching sometimes happens among university system names and university names, such as “univ bordeaux” and “univ Bordeaux 1”. So the postcode limitation is also required for solving the STM problem of French institutions.

Evaluation

In order to evaluate the effectiveness of the rule-based IND method proposed in this study, a test was performed in different fields using two classic indicators from the field of information retrieval: precision and recall. Precise explanations of how we applied these two notions follow below. To guarantee the reliability of the test result, two groups consisting each of two investigators were formed (working independently). First, the rule-based system was applied for each field separately, leading to sets of ISSs each supposed to refer to the same institution (the unique name we choose plays no role here). A threshold of 2 ($T = 2$) was used in step 3 so as to filter out unusual cases. Then it was checked if all institutional names in an ISS really refer to the same institute. Checking was done using the WoS information and several full text databases. If necessary further information was obtained from the Internet, in particular institutional websites. We like to point out here that the workload for checking correctness of a proposed ISS was huge. In some cases institution names looked very similar but actually did not refer to the same institution. An example is the case of univ Bordeaux 1 and univ Bordeaux 2. Clearly a lot of background information is necessary for manual inspection of some subtle differences between two institution names. Existing similarity-based method are not capable of performing correctly, and neither does ours. We think, however, that our approach is better than most.

Precision

The precision ratio is applied on ISS level. This means that it is checked if all institution names in an ISS refer to the same institute. If not the ISS is classified as a mismatch even if there is a partial correct match. Hence precision is determined by considering all ISS found and checking if the names in each ISS refer to the same institute. The announced precision ratios refer to the percentages of correct ISSs.

The final precision ratios for the four fields are shown in Table 3.

Table 3 shows that precision levels in mathematics and computer science are higher than those in psychology and economics and business. It seems that authors in mathematics and computer science are more rigorous in their writing style (especially in the address information) while this is less true for the social sciences. Besides belonging to universities and research agencies, investigators in psychology may also belong to hospitals. Authors engaged in economic and business studies may also take part in business activities organized by other organizations, so that affiliation relationships related to these authors are complicated.

Recall

Theoretically the number of ISSs (different institutions) in each field should be provided in advance to meet the requirement for recall ratio calculation. However, this is not possible in practice. Hence recall ratios for the rule-based IND method were estimated by the results on small data sets. Concretely, the following procedure has been applied: all authors (as defined before, namely surname and initials) that published at least one article were brought together in an author list. Then 30 authors were chosen randomly and all articles (co-)authored by these randomly chosen authors formed the test database. All ISSs were found manually by two investigators. Their result was compared with the result obtained by our automatic rule-based method and a recall ratio was obtained. We performed this test just once for each subject. Hence our results are just indicative. As we have only a relatively small test database we applied Step3 once with $T = 1$ and once with $T = 2$. The resulting lists of obtained unique institutes (ISSs) are denoted by F1 and F2. Recall values are shown in Table 4.

Clearly recall is higher for F1 than for F2. Moreover, the rule-based method performs better, in terms of recall, in mathematics and computer science than in the social sciences. This was also the case for the precision values.

Reasons for mismatching or missing ISSs

Although the experimental result shows that the rule-based IND method performs well in terms of precision, the recall ratios are somewhat disappointing and some institutions

Table 3 Precision test

Subjects	Precision (%)
Mathematics	94.0
Computer science	92.7
Psychology	86.1
Economics and business	84.2

Table 4 Recall test

Subjects	Recall	
	F1 (%)	F2 (%)
Mathematics	87.5	50.0
Computer science	76.5	64.7
Psychology	50.0	10.0
Economics and business	50.0	50.0

remain unrecognized in this study. Analyzing failures in the recall test, we found the following reasons:

1. Low document frequency. Some unusual institution names have a very low document frequency in the WoS, so that it is difficult to be matched with another institution name leading to a successful solution of the STM problem.
2. Failing rule on division matching. In some cases, a real institute may not pass rule6 because an institution's name doesn't share any segment name with other ones.
3. Low word similarity. Although relatively low similarity levels in rule1 or rule2 were required, concretely for the Jaccard similarity we used 0.6 and for the share rule we used the threshold 2, some potential ISSs could not reach the threshold and some names referring to the same institution turned out to be very dissimilar.
4. Test across author blocks. In the test to obtain the recall ratio, some institution names across several author blocks were used by the investigators to check if names refer to the same institute or not. Yet, this is actually out of scope for rules based on author blocks.

Among the four reasons for failure the effect of the first reason is the most significant. This means that a large number of documents is required to obtain sufficient information for a correct operation of our rule-based algorithm.

Conclusion

It was found that there are various formats of names for a single institution, which, moreover may change over time. Besides different writing styles, some STM problems occur because of merging or splitting of institutions. Although the strategy based on institution registration is helpful to standardize institution names (Abramo et al. 2011; D'Angelo et al. 2011; Jiang et al. 2011), it only works in specific and highly controlled circumstances. In order to perform reliable and accurate studies, e.g. for evaluations, on a large scale an automatic IND technique based on data mining may be the only option.

Several earlier studies on AND used word similarity of institution names in an auxiliary way but most did not perform a full evaluation of this aspect. An exception is Jiang et al.'s study (2011) showing that a technique based on NCD for institution name clustering worked well reaching an average precision of 83 %. However, the dataset used in this study consisted mainly of articles from one university and the different formats of institution names were relatively simple, so that the performance test of this NCD-based technique is not convincing. By contrast, the experiment in this study is implemented on a large dataset in four field, mathematics and computer science on the one hand and psychology and economics and business on the other. The lowest precision still reached

84.2 %. The experimental result shows that most names of well-known institutions have been mapped into their institutions and generally the recall ratio is good. Therefore it can be concluded that the IND algorithm proposed in this study works well. Recently a keyword-based method has been proposed by (Morillo et al. 2013) to deal with the question of institution classification. Although they obtained high values of precision and recall, we think that their method depends heavily on the specific application. Because of these special circumstances we do not see it feasible to make an experimental comparison between our method and the few existing attempts.

Because of differences in publications habits between fields, the structure of institution names as used in the scientific literature varies considerably. Such differences were not taken into consideration when IND techniques were applied in previous studies; nevertheless an institution usually is evaluated by discipline, so the application of IND techniques in different fields should be given more attention. As far as the performance of IND technique is concerned in this study in four fields, precision and recall in the hard sciences (mathematics and computer science) are higher than in the soft sciences (psychology and economics and business). One possible reason is that research approaches in the soft sciences are less standardized. We conclude from this observation that one should give more attention to the specific situation of the social sciences, so that higher precision and recall values can be reached.

As summarized before our rule-based IND algorithm performs in an acceptable way in general and the adaptability of the technique in different fields is good as well. Yet, for some cases, such as for general institution names and Institutions with low document frequencies our approach should be improved. Our rule-based approach has been used in a study to answer the question if first rate scientists mainly work at first rate institutions.

Acknowledgments We would like to thank Qiuru Peng, Hui Lin, Xueqin Jiang, and Zengli She from the college of information science and technology for their work on data verification. The authors are supported by Grant No. 13CTQ031 of the National Social Science Fund of China.

References

Abramo, G., Cicero, T., & D'Angelo, C. A. (2011). A field-standardized application of DEA to national-scale research assessment of universities. *Journal of Informetrics, 5*(4), 618–628.

Alias-i. (2002). http://alias-i.com/lingpipe/web/about.html Accessed 13 May 2013.

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2012). Automatic annotation of ambiguous personal names on the web. *Computational Intelligence, 28*(3), 398–425.

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A., & 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.

Csajbók, E., Berhidi, A., Vasas, L., & Schubert, A. (2007). Hirsch-index for countries based on essential science indicators data. *Scientometrics, 73*(1), 91–117.

D'Angelo, C. A., Giuffrida, C., & Abramo, G. (2011). A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. *Journal of the American Society for Information Science and Technology, 62*(2), 257–269.

DeBruin, R. E., & Moed, H. F. (1990). The unification of addresses in scientific publications. In L. Egghe & R. Rousseau (Eds.), *Informetrics 89/90. Selection of papers submitted for the 2nd International Conference on Bibliometrics, Scientometrics and Informetrics* (pp. 65–78). Amsterdam: Elsevier.

Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics. Quantitative methods in library, documentation and information science*. Amsterdam: Elsevier.

French, J. C., Powell, A. L., & Schulman, E. (2000). Using clustering strategies for creating authority files. *Journal of the American Society for Information Science and Technology, 51*(8), 774–786.

Galvez, C., & Moya-Anegón, F. (2006). The unification of institutional addresses applying parametrized finite-state graphs. *Scientometrics, 69*(2), 323–345.

Galvez, C., & Moya-Anegón, F. (2007). Standardizing formats of corporate source data. *Scientometrics, 70*(1), 3–26.

Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles, 37*, 241–272.

Jiang, Y., Zheng, H. T., Wang, X., Lu, B., & Wu, K. (2011). Affiliation disambiguation for constructing semantic digital libraries. *Journal of the American Society for Information Science and Technology, 62*(6), 1029–1041.

Kim, S. W., & Cho, S. Y. (2013). Characteristics of Korean personal names. *Journal of the American Society for Information Science and Technology, 64*(1), 86–95.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*, 707–710.

Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-based bootstrapping for large-scale author disambiguation. *Journal of the American Society for Information Science and Technology, 63*(5), 1030–1047.

Morillo, F., Aparicio, J., González-Albo, B., & Moreno, L. (2013). Towards the automation of address identification. *Scientometrics, 94*(1), 207–224.

Narin, F., Stevens, K., Anderson, J., Collins, P., Irvine, J., Isard, P., et al. (1988). On-line approaches to measuring national scientific output: a cautionary tale. *Science and Public Policy, 15*(3), 153–163.

Onodera, N., Iwasawa, M., Midorikawa, N., Yoshikane, F., Amano, K., Ootani, Y., et al. (2011). A method for eliminating articles by homonymous authors from the large number of articles retrieved by author search. *Journal of the American Society for Information Science and Technology, 62*(4), 677–690.

Pereira, D. A., Ribeiro-Neto, B., Ziviani, N., Laender, A. H. F., & Gonçalves, M. A. (2011). A generic web-based entity resolution framework. *Journal of the American Society for Information Science and Technology, 62*(5), 919–932.

Praal, F., Kosten, J., Calero-Medina, C., & Visser, M. S. (2013). Ranking universities: The challenge of affiliated institutes. *Proceedings of the 18th International Conference on Science and Technology Indicators.* Sept. 4–6, 2013, Berlin, 284–289.

Richardson, G. (2010). Automated country name disambiguation for code set alignment. *Proceedings of the 14th European Conference on Research and advanced technology for digital libraries*. Springer-Verlag Berlin, Heidelberg, 498–501.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology, 43*(1), 1–43.

Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis. *Journal of the American Society for Information Science and Technology, 63*(9), 1820–1833.

Tang, J., Fong, A. C. M., Wang, B., & Zhang, J. (2012). A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering, 24*(6), 975–987.

Taşkın, Z., & Al, U. (2013). Institutional name confusion on citation indexes: The example of the names of Turkish Hospitals. *Procedia—Social and Behavioral Sciences, 73*, 544–550.

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. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics, 62*(1), 133–143.

Yang, K. H, Peng, H. T., & Jiang, J. Y. (2008). Author name disambiguation for citation using topic and web correlation. *Proceedings of the 12th Conference in the series of European Digital Library conferences* (ECDL2008). Sept.19, 2008, Aarhus, 185–196.