Performance of Reference Analysis on Papers in Single Subject Category Journals†‡

XuanZhen Liu* and Hui Fang**

*Nanjing Medical University, Library, Nanjing 210029, China, <lxz@njmu.edu.cn>
** Nanjing University, State Key Laboratory of Analytical Chemistry for Life Science,
School of Electronic Science and Engineering, Nanjing 210023, China, <fanghui@nju.edu.cn>

XuanZhen Liu received a bachelor's degree in medical information science from Jilin University in Jilin, China, in 1994, and a master's degree in library and information science from Nanjing University in Nanjing, China, in 2004. She is now an associate senior librarian at the Nanjing Medical University Library. Her current area of research is bibliometrics.

Hui Fang received a bachelor's degree in radio engineering (1990) and a master's degree in signal processing (1993) from Southeast University in Nanjing, China, and a PhD in electroanalytical chemistry from Nanjing University in Nanjing, China (1998). He is now an associate professor at the School of Electronic Science and Engineering, Nanjing University. His research interests include information processing, data mining, artificial intelligence, instruments and instrumentation, and bibliometrics.

Abstract: Reference analysis is a convenient method for classifying scientific papers into subject categories at publication level. When it is applied to a paper in a single subject category journal, it can recognize the paper's categories other than the journal category. We evaluate the performance of reference analysis with two existing threshold-setting methods for such papers in two physiology journals. The performances of reference analysis with the two threshold-setting methods are also analysed according to the citation distributions of the referenced categories. The numbers of citations to the referenced non-paper categories distribute around a baseline. Introducing a baseline factor into one of the two methods improves the performance of the reference analysis. Errors in the reference analysis come from the various citing behaviours of different authors. Although the two journals used in this study are labelled by the same category, they each have their own focus, which was determined from their topic distributions obtained using the proposed method. This finding matches the author guidelines of the two journals. The distribution of the number of subject categories of each paper is also given.

1.0 Introduction

These days, academic administration offices demand the classification of papers at publication level, which cannot always be satisfied by classification at journal level. For example, the institution of one of the authors counted its publications in each research subfield using the Thomson Reuters Web of Science (WoS) for policy decision-making purposes. It was found that its publications in the subfield of "cardiac and cardiovascular systems" were severely underestimated. A detailed investigation revealed that its many papers in this subfield were published in journals labelled only by the category "physiology" in the WoS. In the WoS, "physiology" includes resources concerned with the normal and pathologic functioning of living cells, tissues, and organisms. It is possible for papers in "cardiac and cardiovascular systems" that focus on the physiology aspect to appear in journals labelled

518

Knowl. Org. 43(2016)No.7
X. Liu and H. Fang. Performance of Reference Analysis on Papers in Single Subject Category Journals

only by "physiology" in the WoS. As a result, such papers are assigned only to the "physiology" category, which ignores their "cardiac and cardiovascular systems" content. This problem also occurs in other subfields.

The classification of scientific papers into appropriate subject fields at publication level, which is one of the basic preconditions of many bibliometric analyses (Glänzel, Schubert and Czerwon 1999; Waltman and Van Eck 2012), can be achieved using a clustering methodology, in addition to the direct use of the subject category at journal level provided by an indexing database (such as the WoS or Elsevier's Scopus). Clustering papers into categories can adopt co-word analysis (Callon et al. 1986), linguistic analysis (Ibekwe-SanJuan et al. 2002), co-citation analysis (Griffith et al. 1974; Small and Sweeney 1985; Small et al. 1985; Klavans and Boyack 2010), combinations of co-citation and co-word analysis (Braam et al. 1991; Su et al. 2010), or citation analysis (Gouvêa Meireles et al. 2014). Believing that direct citations provides a stronger indication of the relatedness of the publications, Waltman and Van Eck (2012) used direct citations to cluster all of the approximately 10 million papers (consisting of articles, letters, and reviews) indexed in the WoS from the period 2001-2010 as an application.

Reference analysis (Small 1987), which also makes use of direct citation relationships, can classify individual papers into existing subject categories. It only needs to download information related to the paper for classification. Reference analysis has been applied to paper classification in multidisciplinary and general journals (Glänzel, Schubert and Czerwon 1999), as well as social science journals (Glänzel et al. 1999), and has been used to improve the SCImago Journal and Country Rank subject classification (Gómez-Núñez et al. 2011). López-Illescas et al. (2009) expanded the paper set of a subfield from papers in specialist journals to include papers in other journals that cite over a certain proportion of its references from this subfield.

Most existing papers on clustering research publications focus on their proposed methods and clustering results, but neglect to validate the correctness of the results. There are only a few papers that verify the correctness of clustering research publications. Joorabchi and Mahdi (2011) tested the precision and recall of reference analysis using 1,000 publications with the help of a group of postgraduate students. Fang (2015) examined the correctness of reference analysis for 114 individual papers published in the *Proceedings of the National Academy of Sciences* (*PNAS*), with the aid of the *PNAS* subject categories. One possible reason for the resistance to testing the correctness of clustering research publications is that it is a tedious and time-consuming manual task. Besides, to ensure the reliability of the correctness test, experts on

the subjects of the papers to be tested are required. In López-Illescas et al. (2009), experts in the field of oncology qualitatively judged the papers to be classified into "oncology," but did not address other subjects to which the papers might belong. The lack of expert checking of the correctness of a publication clustering method publication-by-publication is understandable. Because of the over-competitive research environment (Berezin 2001; Fang 2011), experts do not have much time to help bibliometric researchers determine papers' subject categories.

Here, we apply reference analysis to papers in single subject category journals to show that it can recognize the categories to which a paper belongs (the paper categories) in addition to the journal category in light of the fact that such papers do not necessarily only belong to that journal's category because of knowledge diffusion (Chen et al. 2009). With the help of two physiology experts, we evaluate its performance with two threshold-setting methods, i.e. setting the threshold to some proportion of the total number of references (method 1) or the maximum number times the paper references a subject category (method 2). We also improve the threshold-setting method based on the results.

The remainder of this paper is organized as follows. Section 2.0 describes reference analysis and its threshold-setting methods. Section 3.0 introduces the data and performance measures used in this study. Section 4.0 presents the performance of the reference analysis on papers in single category journals with various threshold-setting methods, analyses this performance according to the distribution of citations among categories the paper referenced (the referenced categories), and proposes a modification to the threshold-setting method as a result. In addition, journal topics are analysed to validate reference analysis with the modified threshold-setting method. Further, we determine the distribution of the number of paper categories in the samples. Finally, Section 5.0 concludes the paper and mentions directions for future research. The limitations of the study are briefly discussed in Section 6.0.

2.0 Methodology

2.1 Reference analysis

Reference analysis categorizes individual papers according to their references (Glänzel, Schubert and Czerwon 1999). This approach is effective in the following main ways (Fang 2015). First, the content of a paper is related to that of its references; the references introduce the background or area of applicability of the paper or relate to the tool or principle adopted by the paper. Thus, the paper has subject categories that are the same as or simi-

lar to those of its references. Second, the intersection of the subject categories of the references can reflect the subject categories of the paper citing them. One reference may cover several aspects that correspond to different subject categories. However, the paper citing that reference relates to a subset of those aspects; they belong to a subset of the subject categories to which the research behind the citing paper corresponds. Suppose a paper belonging to only one subject category has two references: one reference is labelled with subject categories A and B by the database; the other is labelled with A, C, and D. It can then be inferred that the paper belongs to subject category A. The subject categories B, C, and D of the two references may not be related to the subject matter of the citing paper; and reference analysis is able to exclude them from the citing paper's classification, using the correct thresholding tactics which are investigated in this study. Third, this method uses the reliable classification of each reference into the subject categories of publishing journals by experts, i.e. the authors of the references, reviewers, and journal editorial boards. Finally, reference analysis makes use of the existing subject classification system and is immune from the difficulty faced by clustering methods of naming the clustered subject categories (for example, Waltman and Van Eck 2012). As mentioned above, the correctness of reference analysis has been validated in Joorabchi and Mahdi (2011) and Fang (2015).

Fang (2015) expresses reference analysis mathematically. Suppose a paper contains L references (where L is a positive integer). The i-th reference ($i = 1, 2, …, L$) is labelled with n_i subject categories (n_i is also a positive integer). In the WoS, $n_i = 1, 2, …, 6$. The i-th reference is equally likely to be assigned to the n_i subject categories by $1/n_i$ (Waltman, 2012), because it is unclear to which subject category it belongs without further information (Bornmann 2014). Assume the L references are in total labelled with N (a positive integer) subject categories, and matrix $\mathbf{S}_{L \times N}$ represents the assignment of the references to each subject category:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & K & s_{1N} \\ s_{21} & s_{22} & K & s_{2N} \\ M & M & O & M \\ s_{L1} & s_{L2} & K & s_{LN} \end{bmatrix}, (1)$$

where

$$s_{ij} = \begin{cases} 0 & \text{the } i\text{-th reference is not labelled with the } j\text{-th subject category} \\ \dfrac{1}{n_i} & \text{the } i\text{-th reference is labelled with the } j\text{-th subject category} \end{cases}$$

Here, s_{ij} can be regarded as the score of the j-th subject category given to the paper by its i-th reference. The subject categories of each reference are taken to be the subject categories of the journal that publishes it. All the reference information of the inspected paper, such as the journals publishing the references and their subject categories, can be extracted from the full record (including cited references) of the WoS for users to download.

Vector $\mathbf{M}$ is defined as the scores of each subject category, which are obtained from all the references of a paper:

$$\mathbf{M} = (m_1, m_2, …, m_N), (2)$$

where

$$m_j = \sum_{i=1}^{L} s_{ij}, (j = 1, 2, …, N)$$

and can be regarded as the number of times the paper cites the j-th category. The larger m_j is, the more likely the paper belongs to the j-th subject category. Reordering $\mathbf{M}$ in descending order, we have

$$\mathbf{M'} = (m_1', m_2', …, m_N'), (2')$$

Using one of the threshold-setting methods described below, the paper is finally recognized as belonging to the subject categories (called the "recognized categories") that correspond to the first j_R components in $\mathbf{M'}$ (m_1', m_2', …, m'_{j_R}), where j_R is a positive integer satisfying the requirement of the threshold-setting method.

2.2 Example of reference analysis

We use the present paper to illustrate the usage of reference analysis. This paper cites 26 references. One of them is a book that has not been indexed in the WoS, and thus is excluded when classifying this paper. As a result, $L = 25$ for this paper. Here, $N = 9$ because these references are labelled with a total of nine subject categories. The nine referenced categories are represented by SC1, SC2, …, SC9 as follows:

SC1: Multidisciplinary Sciences
SC2: Social Sciences, Interdisciplinary
SC3: Information Science & Library Science
SC4: Computer Science, Information Systems
SC5: History & Philosophy of Science
SC6: Computer Science, Interdisciplinary Applications
SC7: Biology
SC8: Peripheral Vascular Disease
SC9: Radiology, Nuclear Medicine & Medical Imaging

520

Knowl. Org. 43(2016)No.7
X. Liu and H. Fang. Performance of Reference Analysis on Papers in Single Subject Category Journals

Matrix **S** for this paper is then:

$$
\mathbf{S} = \begin{array}{c}
\begin{array}{ccccccccc}
\text{SC1} & \text{SC2} & \text{SC3} & \text{SC4} & \text{SC5} & \text{SC6} & \text{SC7} & \text{SC8} & \text{SC9}
\end{array} \\
\begin{bmatrix}
0.5 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0.5 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0.5 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0.5 \\
0 & 0 & 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0.5 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\end{array}
$$

Accordingly, $\mathbf{M}$ = (0.5, 0.5, 14, 1.5, 2, 4.5, 1, 0.5, 0.5) for this paper. Reordering $\mathbf{M}$ in descending order, we have $\mathbf{M'}$ = (14, 4.5, 2, 1.5, 1, 0.5, 0.5, 0.5, 0.5). Correspondingly, the order of SC1 to SC9 is changed to:

SC1': Information Science & Library Science
SC2': Computer Science, Interdisciplinary Applications
SC3': History & Philosophy of Science
SC4': Computer Science, Information Systems
SC5': Biology
SC6': Multidisciplinary Sciences
SC7': Social Sciences, Interdisciplinary
SC8': Peripheral Vascular Disease
SC9': Radiology, Nuclear Medicine & Medical Imaging

This paper can then be regarded as citing SC1' 14 times, SC2' 4.5 times, SC3' 2 times, SC4' 1.5 times, SC5' 1 time, SC6' 0.5 times, SC7' 0.5 times, SC8' 0.8 times, and SC9' 0.5 times.

This paper discusses the classification of research papers into subject categories. It is hence reasonable for it to cite references in journals labelled with categories "information science and library science" and "computer science," and "interdisciplinary applications." The subject "computer science, information systems" has a close relationship to these two subjects. The references in "history and philosophy of science" cited here discuss research activity and research publications that relate to the topic of this paper. The first reference is published in a journal that is labelled with SC6' and SC7', and it discusses research activities. Some journals in other disciplines, such as "chemistry" and "biology," may also publish a small number of papers discussing research activities. This paper cites two such references, and they are labelled with subject categories SC5', SC8', and SC9'.

2.3 Threshold-setting methods

Threshold-setting method 1—This method selects the recognized categories as the top j_R subject categories that satisfy the following condition: their cumulative percentage of citations exceeds a pre-set threshold. Therefore, it sets threshold P_{Tth} to be the minimum percentage of citations to the recognized categories, as follows:

$$
\frac{\sum_{j=1}^{j_R - 1} m'_j}{\sum_{j=1}^{N} m'_j} < P_{Tth} \le \frac{\sum_{j=1}^{j_R} m'_j}{\sum_{j=1}^{N} m'_j} \, , \quad (3)
$$

where $0 < P_{Tth} < 1$. In other words, P_{Tth} is between the cumulative percentage of citations of the first $j_R - 1$ categories and that of the first j_R categories in $\mathbf{M'}$. According to Eq. (1), $\sum_{j=1}^{N} m'_j = L$. Hence, Eq. (3) can be rewritten as:

$$
\frac{\sum_{j=1}^{j_R - 1} m'_j}{L} < P_{Tth} \le \frac{\sum_{j=1}^{j_R} m'_j}{L} \, . \quad (3')
$$

For example, if P_{Tth} is set as a value in (0, 14/25], then this paper is classified only as SC1' ("information science and library science"), if P_{Tth} is set in (14/25, 18.5/25], then it is classified as SC1' and SC2' ("computer science, interdisciplinary applications"), and so on.

Threshold-setting method 2—This method determines the recognized categories as a collection; the number of citations of each subject category exceeds a pre-set proportion (threshold) of the number of citations of the top recognized category. Therefore, it sets threshold P_{Nth} using the ratio of the citations of a recognized category to the maximum number of citations of a recognized category, as follows:

Knowl. Org. 43(2016)No.7

X. Liu and H. Fang. Performance of Reference Analysis on Papers in Single Subject Category Journals

521

$$m'_{j_R} \geq P_{Nth} m'_1 > m'_{j_R+1}, (4)$$

where $0 < P_{Nth} < 1$. Here, we define the normalized number of citations to a category as m'_j/m'_P for ($i = 1, 2, …, N$). Hence, P_{Nth} satisfies

$$m'_{j_R}/m'_1 \geq P_{Nth} > m'_{j_R+1}/m'_1 . (4')$$

For example, if P_{Nth} is set in (4.5/14, 1], then this paper is classified only as SC1', if P_{Nth} is set in (2/14, 4.5/14], then it is classified as SC1' and SC2', and so on.

3.0 Experiment

Papers (articles and reviews) published in *Acta Physiologica* and the *Journal of General Physiology* in 2014 were used to test the performance of reference analysis with the two threshold-setting methods. The two journals are labelled only by the "physiology" category in the WoS. In this year, there were 141 papers published in *Acta Physiologica* and 83 in the *Journal of General Physiology*. They cite a total of 13,649 references, and of them, there are 13,244 references whose subject categories can be obtained from the WoS (note that of the 144 *Acta Physiologica* items indexed by WoS in 2014, three are not counted here because one contains the annual meeting abstracts and the other two are editorials).

The subject categories of these papers were judged by two physiology experts who read them. Their judgements were based on the definition of individual subject categories by the WoS. This is a time-consuming task. The results of their judgements agree well with each other. Therefore, we regard a paper to belong to a subject category if one expert believes the paper belongs to that category. These subject categories are called the "identified categories." In total, 638 instances of identified categories were obtained by the experts for all the papers. Of these, 13 were recognized by only one of the two experts. In this study, we use the identified categories of a paper as the standard against which the performance of the methods is evaluated. We also define the additional categories as the identified categories that are not the journal categories.

Two types of errors can exist in the paper classification results. One error is that some identified categories may not be recognized. The other is that some recognized categories may not be identified categories. We define the "precision" of the classification method as the fraction of recognized categories that are identified categories, and the "recall" as the fraction of identified categories that are recognized by the method.

4.0 Results and discussion

4.1 Distribution of citations among referenced categories

Both threshold-setting methods determine whether a category is selected to label a paper depending on the relative importance of the referenced categories. In addition, method 1 considers the distribution of the citations to all categories. Figure 1 shows three types of distribution of citations among referenced categories for a single paper. Figure 1(a) is a common type of citation distribution that simultaneously satisfies the condition of the two threshold-setting methods, which are stated in Section 4.2. Figure 1(b) shows the case for a number of papers in which the identified categories, which all have a normalized number of citations larger than 0.28 (for the reason for using this value, see Section 4.2), occupy a much higher proportion of citations. Such a paper requires a higher threshold for method 1. The paper shown in Figure 1(b) has even more identified categories than the others. Figure 1(c) shows another case in which there are many referenced but not identified categories with a normalized number of citations less than 0.28; such a paper requires a lower threshold for method 1. Figure 1(d) shows that although the identified categories of these three papers can be correctly recognized using method 2 (also see Section 4.2), correct recognition for each of the three papers using method 1 requires three separate ranges of P_{Tth}, varying from about 50% to 90%.

4.2 Performance of the reference analysis with the threshold-setting methods

Figure 2 shows the performances of the reference analysis with the two threshold-setting methods. The WoS category "multidisciplinary sciences" was excluded from the results because it provides no useful subject information about the papers. In addition, because the papers used here are published in journals belonging only to the category "physiology," they surely belong to and are thus already assigned to this category. This is a reasonable assumption, because they were classified this way by several experts (their authors, reviewers, and editors) in the process of submission, review, and acceptance for publication. Hence, the subject "physiology" is not counted in Figure 2 and the results below. In other words, the recognized categories and identified categories but not the journal categories are used to determine performance.

In Figure 2(a), the precision of threshold-setting method 1 decreases as the threshold increases, while the recall increases with the threshold. This is because more categories are included in the recognized categories when
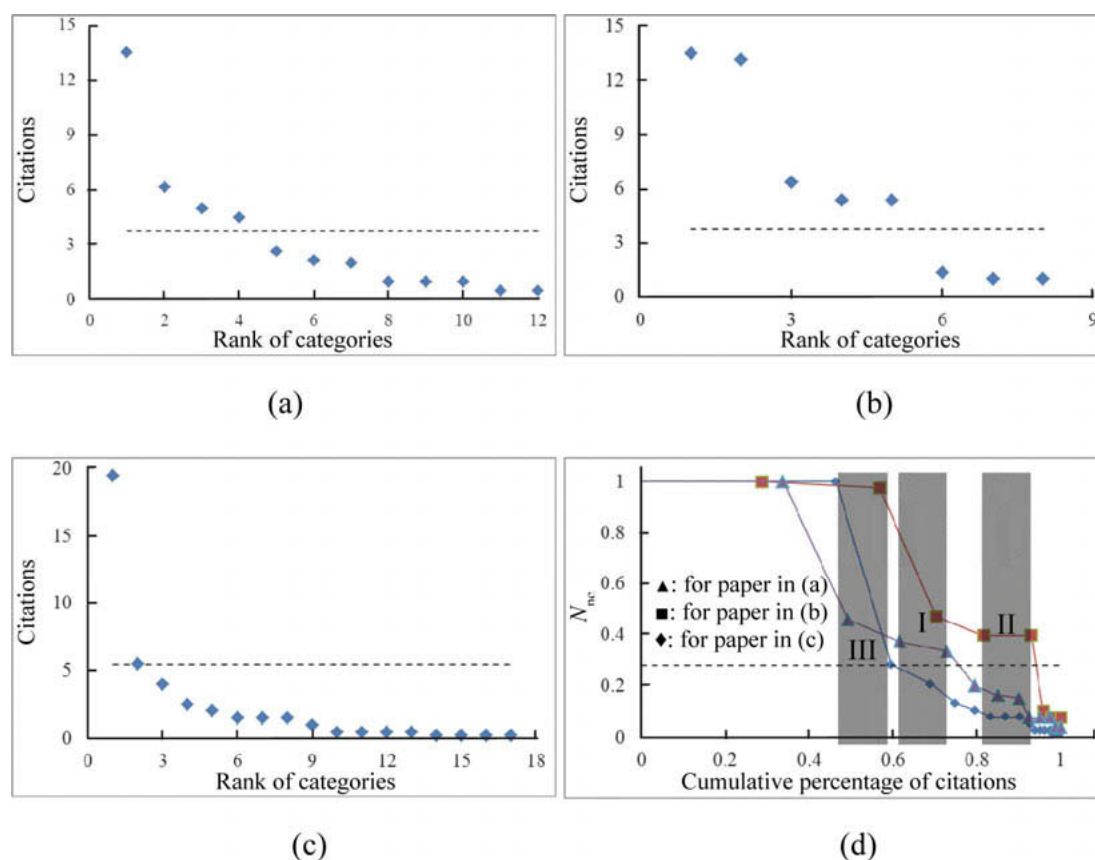
522

Knowl. Org. 43(2016)No.7
X. Liu and H. Fang. Performance of Reference Analysis on Papers in Single Subject Category Journals

Figure 1. Distribution of citations among referenced categories for paper (a) *Acta Physiologica*, 212(3): 239; (b) *Acta Physiologica*, 210(1): 174; and, (c) *Acta Physiologica*, 211(1): 122. (d) Normalized number of citations (N_{nc}: number of citations to a category / the maximum number of citations to a category) versus cumulative percentage of citations to referenced categories for the papers in (a), (b) and (c). Dash line: (0.28 × the maximal number of citations to a category) for (a), (b) and (c), and 0.28 for (d). Section I in (d) is the suitable area to set threshold as the minimum cumulative percentage of citations of recognized categories for paper (a), section II is that for paper (b), and section III is that for paper (c).

the threshold increases. When categories other than the identified categories are included, the precision decreases. When more identified categories are recognized, the recall increases. Similarly, in Figure 2(b), the precision of threshold-setting method 2 increases with the threshold, and the recall decreases as the threshold increases.

As a good classifying method requires both high precision and recall, which change in opposite directions according to the threshold, a compromise must be found. Therefore, we define "correctness" as (precision + recall)/2 to indicate the performance of each method as a whole. In Figure 2, the correctness of both methods reaches its maximum close to the intersection of precision and recall. The correctness of method 2 reaches its maximum of 0.813 at threshold 0.28. The maximal correctness of method 1 is 0.778 at threshold 0.61, but there is a large difference between the precision and recall (0.819 and 0.737, respectively) at this threshold. At a threshold of 0.65, this method achieves a high correctness of 0.775 and a small difference between precision and recall (0.769 and 0.781, respectively). Therefore, we

use $P_{Tth} = 0.65$ for method 1 and $P_{Nth} = 0.28$ for method 2 unless otherwise specified. For example, both methods 1 and 2 classify this paper as SC1' and SC2'.

4.3 Error analysis of reference analysis for threshold-setting methods

In fact, there are differences in citing behaviour among authors (Bornmann and Daniel 2008; Erikson and Erlandson 2014). This leads to some randomness in the distribution of citations among referenced categories, and thus causes errors in the reference analysis. For example, in Figure 3(a), an identified category for a particular paper was not recognized because its numbers of citations satisfied the thresholds of neither method 1 nor 2. On the contrary, in the paper in Figure 3(b), there are two categories recognized by both methods, but neither are identified categories. This paper assessed the effect of PPAR stimulation on cerebral adaptive and therapeutic arterial collateral growth. Of the seven recognized categories, five are identified categories ("peripheral vascular disease," "pharmacology and phar-
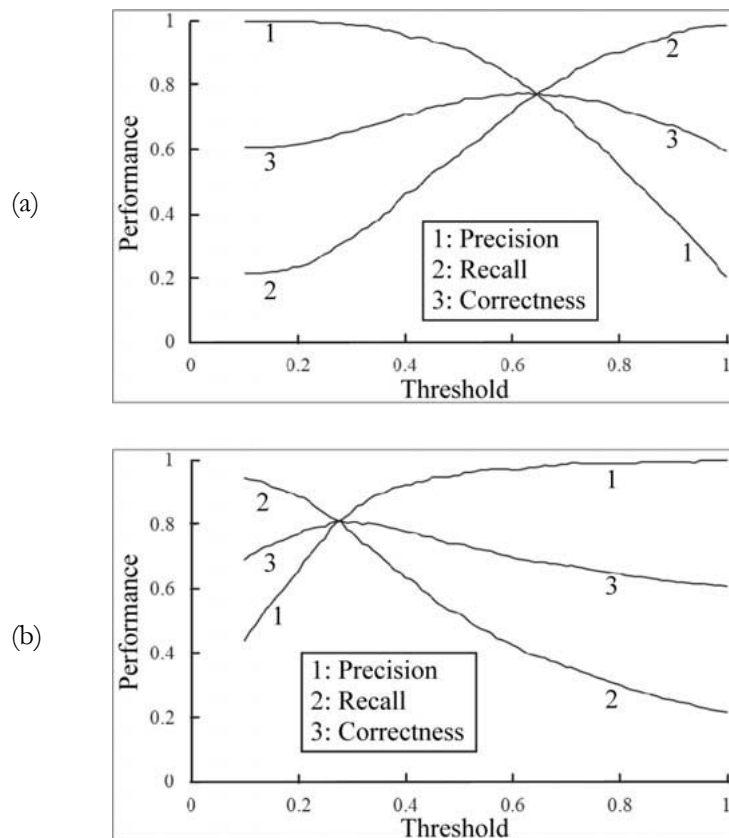
(a)

(b)

Figure 2. Performances of reference analysis with two threshold-set methods. (a) setting the threshold as a certain percentage of the total number of references, and (b) setting the threshold a certain percentage of the maximum number times the paper cites a subject category.

"Precision": the fraction of recognized categories that are identified categories.
"Recall": the fraction of identified categories that are recognized by the method.
"Correctness": ("precision" + "recall")/2.

macy," "cardiac and cardiovascular systems," "physiology," and "cell biology"). The other two ("hematology" and "clinical neurology") relate to the identified categories or the paper itself. "hematology" was included in the recognized categories, because it is one of the categories of the journals publishing some references of this paper. The identified categories are the other categories of these journals. Including "hematology" in the recognized categories can be regarded as a "by-product" of the identified categories. The paper cites references in "clinical neurology" to show its applications, because cerebral arteriogenesis can be used in the therapy of nervous disease such as stroke.

Figure 3(c) and (d) show another type of error for method 2. The maximal citation to a category is not large, so $0.28m'_1$ is lower than the citations to some referenced non-paper categories. For the paper in Figure 3(c), the identified categories are the top six ones. Methods 1 and 2 recognize one and three extra categories, respectively.

4.4 Combination of threshold-setting methods 1 and 2

The analysis above shows that threshold-setting method 2 outperforms method 1 on the whole. However, method 1 can more effectively filter the referenced but not identified categories when the maximal number of citations to a category is not high. In view of this, we tried a combination of methods 1 and 2, namely, the recognized categories satisfying Eqs. (3) and (4) simultaneously. The correctness reaches its peak of 0.814 when $P_{Tth} = 0.78$ for method 1 and $P_{Nth} = 0.28$, which is almost the same as that of method 2 alone.

4.5 Considering a baseline for citation distribution

In the curves in Figures 1 and 3, the citation numbers of categories other than the recognized ones are low and follow a nearly horizontal line. They are introduced into the referenced categories because of their close relation-
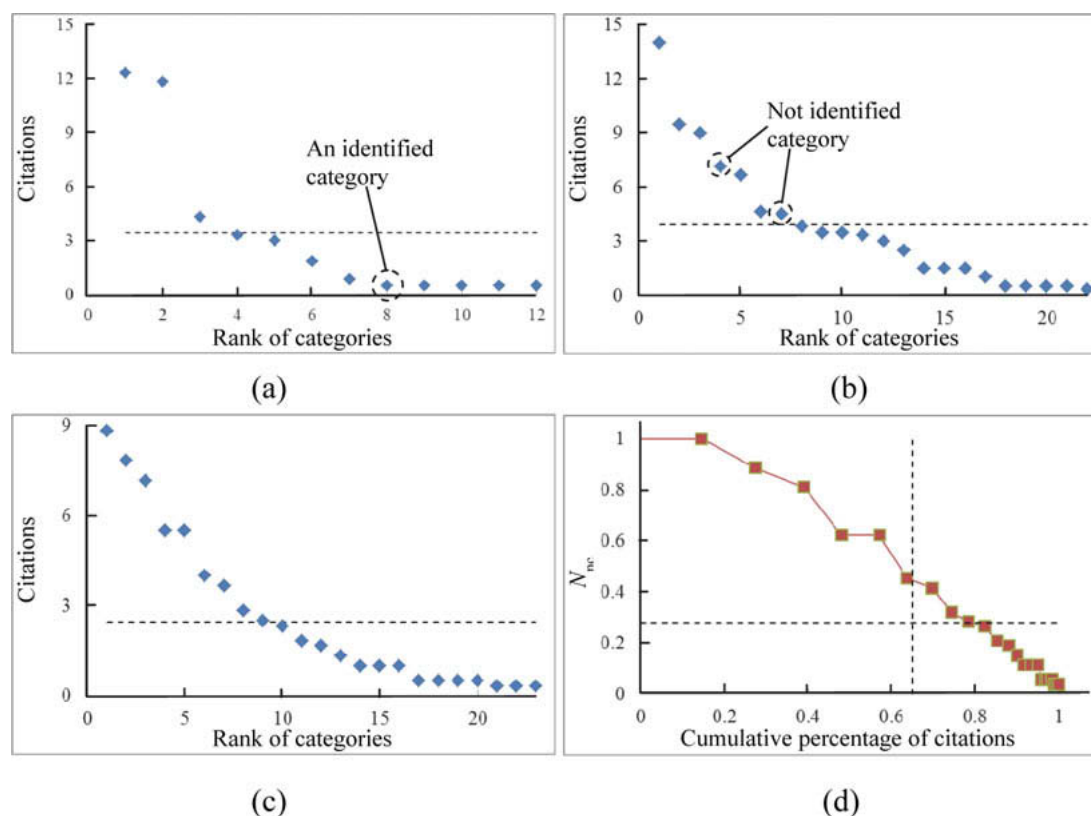
524

Knowl. Org. 43(2016)No.7
X. Liu and H. Fang. Performance of Reference Analysis on Papers in Single Subject Category Journals

Figure 3. Distribution of citations among referenced categories for paper (a) *Acta Physiologica*, 212(3): 191; (b) *Acta Physiologica*, 210(2): 354; and, (c) *Acta Physiologica*, 211(1): 176. (d) Normalized number of citations (N_{nc}: number of citations to a category / the maximum number of citations to a category) versus cumulative percentage of citations to referenced categories for the papers in (c).

"Identified category": the category that experts believe the paper belongs to.
Horizontal dash line: (0.28 × the maximal number of citations to a category) for (a), (b) and (c), and 0.28 for paper (d).
Vertical dash line: cumulative percentage of citations to be 0.65

ships to the paper categories, citing behaviour of author(s), and knowledge diffusion. The number of times a paper cites such categories is with some randomness. Therefore, they can be regarded as a random number in a range around a baseline, which can be regarded as the average number of times the paper cites such categories.

If the maximum number of citations of a paper to a category is not high, such as in the case in Figure 3(c), the threshold of method 2 may fall in the distribution range of the citations to the categories other than identified ones. This causes some extra categories to be included in the recognized categories.

In view of this, method 3 is a modified version of method 2 that introduces a factor m_{bl} that represents the baseline. We then have:

$$m'_{j_R} \geq P_{Nth}(m'_1 - m_{bl}) + m_{bl} > m'_{j_R+1} . (5)$$

However, if $m'_1 < m_{bl}$, then $j_R = 1$.

Applying method 3 to the data in the experiment, the correctness reaches a maximum of 0.835 when $P_{Nth} = 0.17$ and $m_{bl} = 1.77$. Applying this method to the paper shown in Figure 3(c), for which method 2 obtains poorer results than method 1, only one category other than the identified categories is recognized, which is the same result as that of method 1. When applying method 3 to this paper, $P_{Nth}(m'_1 - m_{bl}) + m_{bl} = 3.85$ (with $P_{Nth} = 0.17$ and $m_{bl} = 1.77$), which is between m_2' and m_3'. Therefore, method 3 classifies this paper as SC1' and SC2', just as methods 1 and 2 do.

4.6 Topic analysis of the two journals

Table 1 lists all the recognized categories of papers published on each of the inspected journals in 2014 using method 3. The categories of both journals, "biochemistry and molecular biology," "neurosciences," "cell biology," "cardiac and cardiovascular systems," and "pharmacology and pharmacy" account for a high percentage of the recognized categories of the papers. "Endocrinology and

Recognized categories	*Acta Pysiologica*		the *Journal of General Physiology*	
	N_{p_rec} [a]	Per.	N_{p_rec} [a]	Per.
Endocrinology & Metabolism	56 (1)	13.3%	1 (13)	0.5%
Biochemistry & Molecular Biology	47 (2)	11.2%	66 (1)	30.6%
Neurosciences	47 (2)	11.2%	48 (2)	22.2%
Cell Biology	36 (4)	8.6%	31 (3)	14.4%
Peripheral Vascular Disease	34 (5)	8.1%	2 (9)	0.9%
Cardiac & Cardiovascular Systems	31 (6)	7.4%	5 (6)	2.3%
Sport Sciences	24 (7)	5.7%		
Pharmacology & Pharmacy	23 (8)	5.5%	10 (5)	4.6%
Nutrition & Dietetics	20 (9)	4.8%		
Gastroenterology & Hepatology	11 (10)	2.6%		
Urology & Nephrology	11 (10)	2.6%		
Medicine, Research & Experiment	9 (12)	2.1%	1 (13)	0.5%
Medicine, General & Internal	8 (13)	1.9%		
Pediatrics	7 (14)	1.7%		
Clinical Neurology	6 (15)	1.4%	2 (9)	0.9%
Hematology	6 (15)	1.4%		
Genetics & Heredity	5 (17)	1.2%	1 (13)	0.5%
Behavioral Sciences	4 (18)	1.0%		
Geriatrics & Gerontology	4 (18)	1.0%		
Microbiology			4 (7)	1.9%
Developmental Biology	3 (20)	0.7%		
Immunology	3 (20)	0.7%	1 (13)	0.5%
Obstetrics & Gynecology	3 (20)	0.7%		
Psychiatry	3 (20)	0.7%		
Respiratory System	3 (20)	0.7%		
Physics, Atomic, Molecular & Chemical			3 (8)	1.4%
Agriculture, Dairy & Animal Science	2 (25)	0.5%		
Biology	2 (25)	0.5%	2 (9)	0.9%
Biophysics	2 (25)	0.5%	30 (4)	13.9%
Oncology	2 (25)	0.5%		
Public, Environmental & Occupational Health	2 (25)	0.5%		
Anesthesiology			2 (9)	0.9%
Anatomy & Morphology	1 (30)	0.2%		
Biochemical Research Methods	1 (30)	0.2%		
Critical Care Medicine	1 (30)	0.2%		
Orthopedics	1 (30)	0.2%		
Rehabilitation	1 (30)	0.2%		
Zoology	1 (30)	0.2%		
Chemistry, Medicinal			1 (13)	0.5%
Chemistry, Multidisciplinary			1 (13)	0.5%
Chemistry, Physical			1 (13)	0.5%
Dentistry, Oral Surgery & Medicine			1 (13)	0.5%
Evolutionary Biology			1 (13)	0.5%
Ophthalmology			1 (13)	0.5%
Statistics & Probability			1 (13)	0.5%

Table 1. Recognized categories of papers published in the two inspected journals in 2014.

Note: a. N_{p_rec} represents the number of papers which are recognized as belonging to the category listed in left column. The data in parenthesis following each N_{p_rec} value represents the rank of the category in the whole recognized categories of the journal in terms of N_{p_rec} in descending order.

526

Knowl. Org. 43(2016)No.7
X. Liu and H. Fang. Performance of Reference Analysis on Papers in Single Subject Category Journals

metabolism" ranks first for *Acta Physiologica*, while it is a recognized category for only one paper in the *Journal of General Physiology*. Similarly, "peripheral vascular disease" is a recognized category of 8.1% of papers in *Acta Physiologica*, while for the *Journal of General Physiology*, this proportion is only 0.9%. On the contrary, 13.9% papers in the *Journal of General Physiology* but only 0.5% papers in *Acta Physiologica* are recognized as belonging to "biophysics." Furthermore, there are categories, such as "nutrition and dietetics," "gastroenterology and hepatology," "urology and nephrology," and "pediatrics," which are recognized categories for some papers in *Acta Physiologica*, but not for papers in the *Journal of General Physiology*. In contrast, some physics, chemistry, and mathematics categories emerge in the recognized categories of the *Journal of General Physiology* but not in those of *Acta Physiologica*.

These findings show that, although the two journals both belong to the single category "physiology," *Acta Physiologica* focuses more on clinical medicine while the *Journal of General Physiology* focuses more on mechanism studies. This is partly validated by the scope of the *Journal of General Physiology* (http://jgp.rupress.org/site/misc/ifora.xhtml), which says,

> The *Journal of General Physiology* publishes original work that elucidates basic biological, chemical, or physical mechanisms of broad physiological significance…Although the main emphasis is on physiological problems at the cellular and molecular level, we welcome contributions pertaining to any aspect of general physiology.

In contrast, the scope of *Acta Physiologica* (http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1748-1716/homepage/ForAuthors.html) is not stated as clearly: "*Acta Physiologica*…contains original contributions to physiology and related sciences such as pharmacology and biochemistry, provided the physiological relevance is evident either from the title, the content of the article, or an explanatory statement by the author."

These findings also show that, if the authors of a paper on "endocrinology and metabolism" would like to submit it to journals on "physiology," *Acta Physiologica* is more likely to publish it than the *Journal of General Physiology*. This analysis provides further information to researchers when they select journals to submit papers, especially for those selecting a journal for submitting a paper after rejected by other journals (Neff & Olden 2006; Silberzweig and Khorsandi 2008), as the author guidelines of the target journal may not provide enough detailed information.

4.7 Distribution of the number of additional categories of papers in single category journals

López-Illescas et al. (2009) expanded the paper set of a subject category with papers from additional journals using reference analysis to improve paper retrieval. They found that only considering papers in a subfield's specialist journals leads to an unsatisfactory assessment of research groups. With respect to a certain category, classifying papers at journal level misses some papers that belong to it. At the same time, classifying papers at journal level possibly misses paper categories when labelling papers. For example, paper (*Acta Physiologica* 212(3): 214) is identified as belonging to "physiology," "neurosciences," and "sport sciences." When retrieving papers on "neurosciences" from the current indexing database, it will be missed. However, when retrieving papers on "sport sciences," it will also be missed.

Figure 4 shows the distribution of the number of additional categories of each paper used in this study. Every paper in the sample has at least one additional category. This indicates that category "physiology" has a close relationship with other categories. Papers with three additional categories account for the highest proportion of papers. There are only a few papers with more than five additional categories. It seems that an upper limit of six subject categories is suitable for assigning papers in such journals. This figure was found to be four in the reference analysis of *PNAS* (Fang 2015).
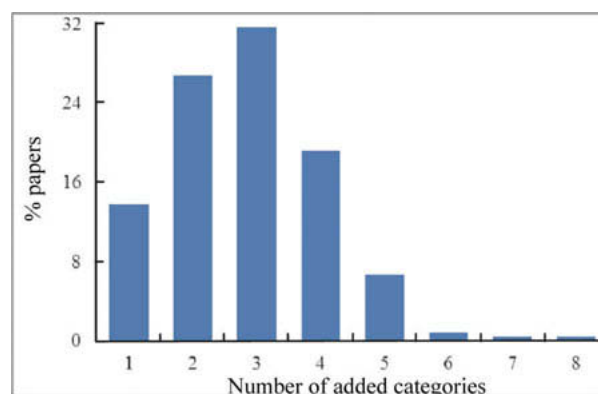

Figure 4. Distribution of number of additional categories of the papers in single category journals used in the experiment.

Additional categories: Identified categories but not the journal category.

5.0 Conclusions

Correctly classifying research papers is important for information retrieval and investigations on research activity, and thus it can provide useful information for academic

administration offices. For example, if the WoS has embedded reference analysis to further classify research papers at publication level, the dilemma mentioned at the beginning of the introduction can be avoided when one searches for academic papers using subject categories provided by the indexing database. Using reference analysis to classify academic papers does not require a large amount of data to be downloaded, and individual papers can be classified into an existing subject category system at publication level. Besides, the results of the reference analysis of a paper can be considered time-independent because they may be measured at the time the paper is published and do not subsequently change. Finally, as shown in the example of this paper, reference analysis has a simple implementation and thus has lower system requirements than other paper classification methods, especially for ordinary users.

Aiming to optimize reference analysis, this work investigates its performance on papers in single subject category journals with different threshold-setting methods, using two journals labelled with only a single, identical category. In this experiment, the threshold-setting method that uses the maximum citations to a category (method 2) performed better as a whole than the method related to the total number of references (method 1). To explain this phenomenon, we inspected the citation distribution among the referenced categories.

Method 1 not only relies on the relative importance of categories but also on the number of referenced categories and paper categories. When there are more paper categories, a higher threshold is required. More referenced non-paper categories require a lower threshold. In comparison, method 2 only relies on the relative importance between categories, and thus there is less uncertainty when determining the threshold. However, when a paper's number of citations to its most cited category is low, method 2 recognizes too many categories because of the random distributions of the citations of referenced non-paper categories around a baseline.

We have tried two modifications of the threshold-setting method. One is a combination of methods 1 and 2. It performs slightly better than method 2. Considering the distribution of citations among referenced non-paper categories, the other modification introduces a factor representing the baseline in method 2 to form method 3. It performs better than methods 1 and 2, achieving up to 83.5% correctness.

The errors in reference analysis come from the various citing behaviours of authors. As a result, the methods recognize some categories to which the paper does not belong, or miss some paper categories. Fortunately, the incorrectly recognized categories relate to the paper or the paper categories.

This work is the first to focus on the application of reference analysis to papers in single subject category journals, and it concludes that threshold-setting method 2 outperforms threshold-setting method 1. Using error analysis, the threshold-setting method 2 is further improved. Further, in this work, the comparisons and improvements are all based on calculating the correctness of the methods. A similar work that also tests the performance of reference analysis using threshold-setting method 2 is Fang (2015), which shows that this method can perfectly classify 78 papers and acceptably classify 25 papers from a total of 114 papers in *PNAS*. This is similar to the results of this work for threshold-setting method 2.

The topics of the two journals used here were analysed by collecting the recognized categories of the papers in each journal. Although both *Acta Physiologica* and the *Journal of General Physiology* are labelled by only one category, "physiology," the former focuses more on clinical medicine and the latter focuses more on mechanism studies. The results agree with the author guidelines of the two journals, as a further validation of the methodology, and might help authors when selecting journals for paper submission.

Therefore, this study suggests incorporating reference analysis in an indexing database that has an existing subject classification system. In addition to improving information retrieval performance, it can 1) assist the system to correctly assign already-published papers to the most relevant subject; 2) detect differences between journal subject categories and paper subject categories; 3) apprise prospective authors of these discrepancies and re-orient them towards the most suitable journal to which their paper should be submitted; and, 4) assist a journal editorial board to predict the categories to which currently submitted papers should be classified, and thus may automatically filter papers that are out of the scope of the journal. The investigation of threshold-setting methods in this work can improve the precision of these applications of reference analysis.

We finally count the number of additional categories of each paper used in this work. An upper limit of six subject categories was found suitable for assigning to these papers. This number provides useful information for determining suitable upper limit of subject categories for classifying papers at publication level in the future.

Glänzel, Schubert and Czerwon (1999) proposed an iterative process that distinguishes the recognized categories of a paper from the recognized categories of its references. We found it did not improve the correctness in our case (these experiments are not included in results). The potential reason for this might come from the references published in the journal that are labelled with several (2-6) categories. For example, a paper published in a journal la-

528

Knowl. Org. 43(2016)No.7

X. Liu and H. Fang. Performance of Reference Analysis on Papers in Single Subject Category Journals

belled with three categories may only belong to two of them. Therefore, correctly recognizing the categories of references on such journals might improve the performance of iterative reference analysis, and this will be tried in future studies to determine whether the performance of reference analysis can be further improved.

6.0 Limitations of the study

Our findings should be generalized with caution. Because of the time-consuming work required to judge the identified categories of each sample paper by experts, our work has focused only on papers in two physiology journals. The results are therefore not transferable to other research fields. Further extensive investigations, using papers in the journals of other research fields, are required to test the universality of our findings.

References

Berezin, Alexander A. 2001. "Discouragement of Innovation by Overcompetitive Research Funding." *Interdisciplinary Science Reviews* 26:97–102.

Bornmann, Lutz. 2014. "Assigning Publications to Multiple Subject Categories for Bibliometric Analysis, an Empirical Case Study Based on Percentiles." *Journal of Documentation* 70:52–61.

Bornmann, Lutz and Hans-Dieter Daniel. 2008. "What Do Citation Counts Measure? A Review of Studies on Citing Behaviour." *Journal of Documentation* 64:45–80.

Braam, Robert R., Henk F. Moed and Anthony F. J. van Raan. 1991. "Mapping of Science by Combined Co-Citation and Co-Word Analysis I. Structural Aspects." *Journal of the American Society for Information Science* 42:233–51.

Callon, Michel, John Law and Arie Rip. 1986. "Qualitative Scientometrics." In *Mapping the Dynamics of Science and Technology*, ed. M. Callon, J. Law and A. Rip. The Macmillan Press Ltd., 103.

Chen, Chaomei, Yue Chen, Mark Horowitz, Haiyan Hou, Zeyuan Liu and Donald Pellegrino. 2009. "Towards an Explanatory and Computational Theory of Scientific Discovery." *Journal of Informetrics* 3:191–209.

Erikson, Martin G and Peter Erlandson. 2014. "A Taxonomy of Motives to Cite." *Social Studies of Science* 44:625–37.

Fang, Hui. 2011. "Peer Review and Over-Competitive Research Funding Fostering Mainstream Opinion to Monopoly." *Scientometrics* 87:293–301.

Fang, Hui. 2015. "Classifying Research Articles in Multidisciplinary Sciences Journals into Subject Categories." *Knowledge Organization* 42:139–53.

Glänzel, W., A. Schubert and H. J. Czerwon. 1999a. "An Item-by-Item Subject Classification of Papers Published in Multidisciplinary and General Journals Using Reference Analysis." *Scientometrics* 44:427–39.

Glänzel, W., A. Schubert, U. Schoepflin and H. J. Czerwon. 1999b. "An Item-By-Item Subject Classification of Papers Published in Journals Covered by the SSCI Database Using Reference Analysis." *Scientometrics* 46:431–41.

Gómez-Núñez, Antonio J., Benjamín Vargas-Quesada, Félix de Moya-Anegón and Wolfgang Glänzel. 2011. "Improving SCImago Journal & Country Rank (SJR) Subject Classification through Reference Analysis." *Scientometrics* 89:741–58.

Gouvêa Meireles, Magali Rezende, Beatriz Valadares Cendón and Paulo Eduardo Maciel de Almeida. 2014. "Bibliometric Knowledge Organization: A Domain Analytic Method Using Artificial Neural Networks." *Knowledge Organization* 41:145–59.

Griffith, Belver C., Henry. G. Small, Judith A. Stonehill and Sandra Dey. 1974. "The Structure of Scientific Literatures. II: Toward a Macro-and Microstructure for Science." *Science Studies* 4:339–65.

Ibekwe-SanJuan, Fidelia and Eric SanJuan. 2002. "From Term Variants to Research Topics." *Knowledge Organization* 29:181–97.

Joorabchi, Arash and Abdulhussain E. Mahdi. 2011. "An Unsupervised Approach to Automatic Classification of Scientific Literature Utilizing Bibliographic Metadata." *Journal of Information Science* 37:499–514.

Klavans, Richard and Kevin W. Boyack. 2010. "Toward an Objective, Reliable and Accurate Method for Measuring Research Leadership." *Scientometrics* 82:539–53.

López-Illescas, Carmen, Ed C. M. Noyons, Martijn S. Visser, Félix De Moya-Anegón and Henk F. Moed. 2009. "Expansion of Scientific Journal Categories Using Reference Analysis: How Can It be Done and Does It Make a Difference?" *Scientometrics* 79:473–90.

Neff, Byan D. and, Julian D. Olden. 2006. "Is Peer Review a Game of Chance?" *Bioscience* 56:333–40.

Silberzweig, James E. and Azita S. Khorsandi. 2008. "Outcomes of Rejected Journal of Vascular and Interventional Radiology Manuscripts." *Journal of Vascular and Interventional Radiology* 19:1620–3.

Small, H. 1987. "The Significance of Bibliographic References." *Scientometrics* 12:339–41.

Small, H. and E. Sweeney. 1985. "Clustering the Science Citation Index Using Co-Citations. I. A Comparison of Methods." *Scientometrics* 7:391–409.

Small, H., E. Sweeney and E. Greenlee. 1985. "Clustering the Science Citation Index Using Co-Citations. II. Mapping Science." *Scientometrics* 8:321–40.

Su, Yu-Min, Ping-Yu Khorsandi and Ning-Yao Pai. 2010. "An Approach to Discover and Recommend Cross-

Knowl. Org. 43(2016)No.7

529

X. Liu and H. Fang. Performance of Reference Analysis on Papers in Single Subject Category Journals

Domain Bridge-Keywords in Document Banks." *Electronic Library* 28:669–87.

Waltman, Ludo. 2012. "An Empirical Analysis of the Use of Alphabetical Authorship in Scientific Publishing." *Journal of Informetrics* 6:700–11.

Waltman, Ludo and Nees Jan Van Eck. 2012. "A New Methodology for Constructing a Publication-Level Classification System of Science." *Journal of the American Society for Information Science and Technology* 63:2378–92.