Applied Intelligence

Some bibliometric procedures for analyzing and evaluating research fields
--Manuscript Draft--

Manuscript Number:	APIN-D-17-00658R2
Full Title:	Some bibliometric procedures for analyzing and evaluating research fields
Article Type:	Original Submission
Keywords:	scientific knowledge discovery; Bibliometrics; science mapping
Corresponding Author:	Enrique Herrera-Viedma, Ph.D. University of Granada granada, SPAIN
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Granada
Corresponding Author's Secondary Institution:	
First Author:	Maria Gutierrez-Salcedo
First Author Secondary Information:	
Order of Authors:	Maria Gutierrez-Salcedo
	M. Angeles Martinez
	Jose Antonio Moral-Muñoz
	Enrique Herrera-Viedma, Ph.D.
	Cristobal Torres-Albero
	Manuel Jesus Cobo
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	Nowadays, measure the quality and quantity of the scientific production is an important necessity since almost every research assessment decision depends, to a great extent, upon the scientific merits of the involved researchers. To do that, many different indicators have been proposed in the literature. In addition, the highly cited papers of a discipline could be discovered by means of citation classics analysis. On the other hand, the extraction of knowledge from the intellectual, social or conceptual structure of a research field could be done by means of science mapping analysis based on bibliographic networks. In this paper, we introduce some of the most important techniques and software tools to analyze the impact of a research field and its scientific structures. Particularly, four bibliometric indices (h, g, hg and q2), the h-classics approach to identify the classics papers of a research field and three science free science mapping software tools (CitNetExplorer, SciMAT and VOSViewer) are shown.

∫Response to the Reviewers to the Report on APIN-D-17-00658R1

Title of the paper: Some bibliometric procedures for analyzing and evaluating research fields

----

GLOBAL RESPONSE

Following referees' suggestions, the main changes done in the manuscript are:

- The initial authors have been fixed.
- The manuscript has been revised according to all referee comments.

- The language style has been improved.

- --

Reviewer #1: All changes have been addressed. Accept as it is.

Answer: Thank you for your comments.

Reviewer #3: I am ok with the revision

Answer: Thank you for your comments.!!!!

Reviewer #4: The authors have addressed my concerns, and as far as I
can tell, also those of the other reviewers. While there are still
some minor language glitches, they should not affect the reader's
ability to understand the material.

Answer: We have checked again the English style of the paper. Thank
you for your comments.!!!

Some bibliometric procedures for analyzing and evaluating research fields

M. Gutiérrez-Salcedo · M. Ángeles Martínez · J.A. Moral-Munoz · E. Herrera-Viedma · M.J. Cobo

Abstract Nowadays, measure the quality and quantity of the scientific production is an important necessity since almost every research assessment decision depends, to a great extent, upon the scientific merits of the involved researchers. To do that, many different indicators have been proposed in the literature. Two main bibliometric procedures to explore a research field have been defined: performance analysis and science mapping. On the one hand, performance analysis aims at evaluating groups of scientific actors (countries, universities, departments, researchers) and the impact of their activity on the basis of bibliographic data. On the other hand, the extraction of knowledge from the intellectual, social or conceptual structure of a research field could be done by means of science mapping analysis based on bibliographic networks. In this paper, we introduce some of the most important techniques and software tools to analyze the impact of a research field and its scientific structures. Particularly, four bibliometric indices (h, g, hg and q^2), the h-classics approach

M. Gutiérrez-Salcedo
Department Management and Marketing, University of Jaén, Jaén, Spain.
E-mail: msalcedo@ujaen.es

M. Ángeles Martínez
Department of Social Work and Social Services, University of Granada, Granada, Spain.
E-mail: mundodesilencio@ugr.es

J.A. Moral-Munoz
Department Nursing and Physiotherapy, University of Cádiz, Cádiz, Spain, Spain.
Institute of Research and Innovation in Biomedical Sciences of the Province of Cadiz (IN-iBICA), University of Cádiz, Cádiz, Spain.
E-mail: joseantonio.moral@uca.es

E. Herrera-Viedma
Department of Computer Science and A.I., University of Granada, Granada, Spain.
E-mail: viedma@decsai.ugr.es

M.J. Cobo
Department of Computer Science, University of Cádiz, Algeciras, Spain.
E-mail: manueljesus.cobo@uca.es

to identify the classic papers of a research field and three free science mapping software tools (CitNetExplorer, SciMAT and VOSViewer) are shown.

Keywords Bibliometrics · h-index · science mapping · citations

1 Introduction

Nowadays there is a huge amount of scientific documents publishing in a great variety of research areas. Take advantage of this information in order to implement specific intelligent information systems, such finding new trends or identifying groups of related scientists, need for special techniques [1][22][36][57]. In this sense, Bibliometrics is an academic science whose aim is to evaluate the research developed by any scientific community in any field. Concretely, Bibliometrics is a set of methods used to study or measure the research through the scientific publications stored or indexed in big bibliographic databases. Many scientific communities use bibliometric methods to explore the impact of their field, the impact of a set of researchers, or the impact of a particular paper [53]. In this sense, Bibliometrics contributes to the progress of science because it allows us to discover information in many different ways: allowing assessing progress made, identifying the most reliable sources of scientific publication, laying the academic foundation for the evaluation of new developments, identifying major scientific actors, developing bibliometric indices to assess academic output, and so on. Therefore, Bibliometrics has become an essential tool in most scientific areas that aims to progress (medicine, mathematics, economics, computer science, physics, sociology, psychology, etc.) [44].

Bibliometrics is mainly focused on to quantify the scientific production and measure its quality and scientific impact [15]. In addition, it deals with the understanding of the social, intellectual and conceptual structure through bibliographic networks [4] (e.g. co-words, co-citation or co-authors, among others).

On the one hand, the production and quality could be measured by means of bibliometric indicators or bibliometric indices. Moreover, performing a citation classics analysis, those papers with highest citation rate could be addressed. On the other hand, bibliographic networks could be analyzed by means of science mapping analysis.

In this paper, we present a review of bibliometric techniques to measure the impact of a set of papers and a description of three science mapping software tools. In this sense, four bibliometric indices (h, g, hg, and q^2) are described [35][25][3][10]. Moreover, an approach to identify the classic papers of a scientific research field based on the h-index [45] is showed. Finally, we introduce three free science mapping software tools: CitNetExplorer [24], SciMAT [17] and VOSviewer [23].

This paper is organized as follows: in Section 2 bibliometrics and science mapping analysis are briefly introduced. Section 3 presents four bibliometric indices. In Section 4 the H-Classics is shown. Section 5 presents the science

mapping analysis software tools. Finally, in Section 6 some conclusion are drawn.

2 Bibliometrics and Science Mapping Analysis

Formally, Bibliometrics are a set of methods, which can be used to analyze academic literature quantitatively and its changes over time [15]. It is devoted to assess and analyze the academic research carried out in different countries, universities, research centers, research groups and journals. It could be used as an objective criterion to evaluate the research developed by scientists, and therefore, it is increasingly valued as a tool for assessing scholarly quality and productivity [46].

In bibliometrics, there are two main methods for exploring a research field: performance analysis and science mapping [50][53]. While performance analysis aims to evaluate the citation impact of the scientific production of different scientific actors, science mapping aims to display the conceptual, social or intellectual structure of scientific research, as well as its evolution and dynamical aspects.

Performance bibliometric analysis is focused on the measurement of scientific output through quality and quantity indicators. In the literature, there are a great variety of bibliometric indicators to measure the scientific output of researchers:

- **Production indicators**: total number of published papers, number of papers published in a period of time, etc. They are focused on the counting of the papers published by the different scientific actors.
- **Impact indicators based on received citations**: total number of citations (including or excluding self citations), average number of citations per paper, self citations rate, number of highly cited papers, h-index [35], g-index [25], hg-index [3], q^2-index [10], crown-index [46], etc.
- **Indicators based on the impact of the journal**: Impact Factor [28][30], SJR [56], normalized impact factor, relative citations ratio [38], etc.

On the other hand, science mapping or bibliometric mapping is a spatial representation of how disciplines, fields, specialties, and documents or authors are related to one another [60]. It has been widely used to show and uncover the hidden key elements (documents, authors, institutions, topics, etc.) in different research fields [12][13][14][18][37][43][48][49][54][64].

In science mapping analysis, the scientific knowledge could be understood as a complex system, where the network structure is often used to model the interaction among scientific actors (authors, journals, keywords, references, etc.) [6]. Depending on the kind of aspects that will be represented, three kinds of bibliographic or bibliometric networks [4] could be identified:

- **Collaboration networks** are used to show how authors or institutions relate to others in the field of scientific research. The most common kind of

collaboration network is co-author network. With this type of network can be discovered, for example, groups of regular authors, influence authors, hidden communities of authors, relevant institutions in a specific research field, etc. [34].

– **Conceptual networks** [11][15] represent relations between concepts or words in a set of publications. That means that, for instance, words which appear together in a document, will be related in a network. It is also known as the co-words network. This type could be used to understand the topics covered by a research field to define what are the most important and the most recent issues. It could also help in the study of the evolution of subjects over time and it could give a good impression of cognitive relations between different research groups.

– **Publication Citation networks** [42][59] show relationships between nodes which represent publications, while the edges can have different interpretations depending on the network type (co-citation, bibliographic coupling or direct link).

Finally, it should be noted that science mapping analysis involves a specific workflow [7][16]:

1. **Data retrieval** from bibliographic databases, such as ISI Web of Science (ISIWoS), Scopus or Google Scholar.
2. **Data preprocessing** in order to improve the quality of the data, clean the errors and complete the missing fields. Different preprocessing methods can be applied, among which it is worth mentioning those that detect duplicate and misspelled items, time slicing, and data reduction.
3. **Unit of analysis selection**: journals, documents, authors, affiliations, words or references, among others.
4. Once the unit of analysis has been selected, several relationships among them can be established (i.e. co-occurrence, coupling or direct linkage) in order to **build the bibliometric network** [4].
5. When the network of relationships between the selected units of analysis has been built, a **normalization process** is needed to correct the data for differences in the number of occurrences [65].
6. Different techniques could be applied to **build science maps**, such as principal component analysis or clustering algorithms, among others [7].
7. Several **analysis** techniques could be applied to yield useful knowledge [16]: network analysis [19], longitudinal analysis [32], performance analysis, etc.
8. The networks and results obtained should be shown up using a quality **visualization** techniques in order to achieve a good understanding and better interpretation of the output [16].
9. At the end of the process, the analysts have to **interpret the results** and maps using their experience and knowledge. In the interpretation step, the analyst looks to discover and extract useful knowledge that could be used to make decisions.

3 Bibliometric indices

Measuring the scientific output of researchers is an important task for the scientific community. Furthermore, most of the scientific policy making depends on the scientific merits of the involved researchers [2]. Nowadays, this evaluation process is performed by means of the computation of bibliometric measures due to the benefits of obtaining an unbiased and fair criterion [58].

In this sense, the h-index and some of its variants have received a great attention from the scientific community due to the easiness of computation and its balance between quantity of publications and their impact [3]. In fact, there exists a great variety of h-index based bibliometric indices. So, in this Section we focus on the main basic indices (h-index [35] and g-index [25]) and some indices based on aggregations (hg-index [3] and q^2-index [10]).

The **h-index** [35] was proposed and defined by the physicist Hirsch in order to measure the productivity of a researcher taking into account his/her publications and citations achieved. It combines in a bibliometric indicator, the amount of publications and their impact, so it allows to simplify the characterization of the researcher's scientific output. Despite its criticism, h-index is nowadays one of the most important impact measures of the researcher's oeuvre. Hirsch defined the h-index as follows [35]:

> "A scientist has index h if h of his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have $\leq h$ citations each."

The computation of the h-index of a researcher is simple, and in fact, this is one of its most important advantages. According to Hirsch [35], the first step is to order the publications list by citations count in decreasing order. The second step is to find the point where the number of citations crosses the publication order, which corresponds with the h-index. To illustrate the h-index, Figure 1 shows an example of the h-index from a set of 100 papers (citations have been computed randomly). The blue shape represents the citations achieved by each paper in decreasing order. The brown line represents the h-index at each point. The cross between both stands for the h-index. That is, the first document with a citations count greater or equal to h.

Moreover, the h-index presents some advantages [2], such as, objectivity [20] and robustness [66] due to it is insensitive to a set of lowly cited paper. On the other hand, it presents some drawbacks [2]: differences among disciplines [35] or biased by the length of the scientific career (citations increase over time).

Furthermore, Egghe argued that the h-index is also insensitive to highly cited papers due to once a paper is selected to belong to the top h papers, it is no longer used in the determination of h. That is, the h-index computed in subsequent years will not be influenced by the citations achieved by the highly cited papers that just belong to the top h papers [25].

To overcome this drawback and keep up the main advantages of the h-index, Egghe proposed the **g-index** [25] which give more weight to the highly cited papers. Its computation is quite similar to the h-index, starting from a
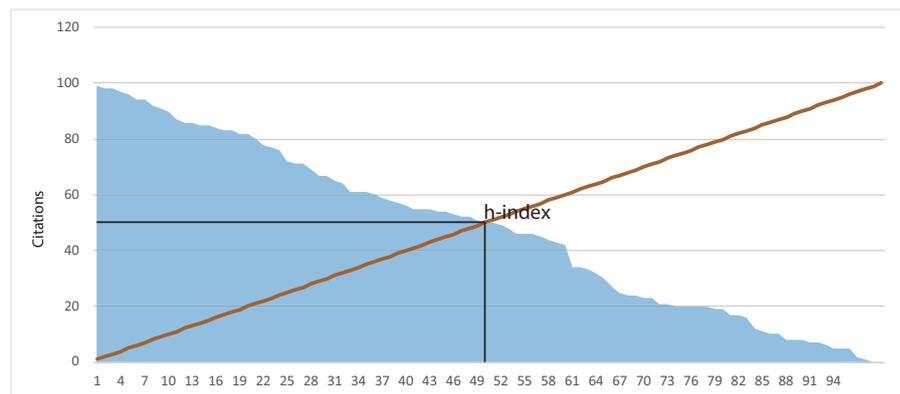
Fig. 1 Example of h-index.

set of papers in decreasing order of citations. Egghe defined the g-index as follows:

> *"A set of papers has a g-index g if g is the highest rank such that the top g papers have, together, at least g^2 citations. This also means that the top $g + 1$ papers have less than $(g + 1)^2$ papers."*

From its definition is clear that $g \geq h$, since h satisfies the requirement that the top h papers have at least h^2 papers and since g is the largest number with this property.

Although the g-index was developed to overcome some disadvantages of the h-index and is successful in evaluating the production of a researcher incorporating the actual citations of his/her papers, it also presents some drawbacks that must to be taken into consideration. Particularly, the g-index is biased by a very successful paper [3]. For example, a researcher with 10 publications (one with 20 citations and the remaining with 0 citations) will have a g-index bigger than a researcher with 10 publications, each one with 5 citations.

Both, h-index and g-index could be viewed as complementary. According to Rousseau [55]: *"As to the h- and the g-index they do measure different aspects of a scientists publication list. Certainly, the h-index does not tell the full story, and although a more sensitive indicator than the h-index, neither does the g-index. Taken together, g and h present a concise picture of a scientists achievements in terms of publications and citations"*.

Following the arguments of Rousseau, Alonso et al. proposed a bibliometric index called **hg-index** which combined both indices. The hg-index is defined as follows [3]:

> *"The hg-index of a researcher is computed as the geometric mean of his h- and g-indices, that is: $hg = \sqrt{h \times g}$"*

It could be demonstrated that $h \leq hg \leq g$ and $hg - h \leq g - h$, which means that the hg-index is a value closer to h than g [3]. This property avoids the

problem of the big influence introduced by a highly cited paper in the g-index, since it is a penalty of the g-index in the case of a very low h-index [3].

The three indices described above (h-index, g-index and hg-index) allow to identify the group of high-performance papers with respect to the researcher's career [41]. In this sense, Rousseau introduced the concept of h-core as the set of papers that belong to the h-index [55]. On the other hand, other bibliometric indices characterize the impact of the papers inside the core [8][26][40][41].

According to Bornmann, both types of indices represent different dimensions of the researcher output, but in fact, they are complementary [8]. Thus, both types of indices could be combined using an aggregation operator.

In this sense, Cabrerizo et al. proposed the $\mathbf{q^2}$-**index** [10] which is computed as the geometric mean of the h- and m-indices: $q^2 = \sqrt{h \times m}$. The m-index [8] is defined as the median number of citations received by the papers in the h-core. That is, the number that separates the higher half of a set from the lower half.

As in the case of the hg-index, there is an arithmetic relationship between the h-index and q^2-index, that stands that $h \leq q^2 \leq m$ and that $q^2 - h \leq m - q^2$. It indicates that the q^2-index is nearer to h than m. In the case of very low h-index, the m-index is penalized.

The q^2-index presents some advantages, such as, the simplicity of computation, more granularity than the h-index and to take into account both quantitative and qualitative dimensions [10].

Finally, a global example and comparison between the four described indices are shown in Figure 2. The blue shape represents the citations achieved by each document in decreasing order. Similarly, the green shape stands for the cumulative citations. That is, the value of the first document (that with the highest citation) stands for $citations_1$, the value of the second document stands for $citations_1 + citations_2$, and so on. Moreover, the brown and red lines represent the h and g curves, respectively. As in Figure 1, the intersection between the blue shape and brown line represents the h-index (with a value equal to 12). Similarly, the intersection between the red line and green shape stands for the g-index (with a value equal to 17). We should point out that in this case, the intersection occurs at the middle of two documents, so the g-index is the first document where the cumulative citations are lower or equal to g^2 (g-index is an integer value). Furthermore, in Figure 2 the values of the aggregated indices are shown. Thus, the hg-index is placed between the h and g indices, as it is pointed in its definition, with a value of 14.28. Similarly, the q^2-index is shown at the right of the hg-index. Also, the m-index is shown, which stands as the median of citations received by the h-core. By contrast, the m-index represents a number of citations (21 in the example), where the remaining indices explained represent a number of documents that meet the threshold.
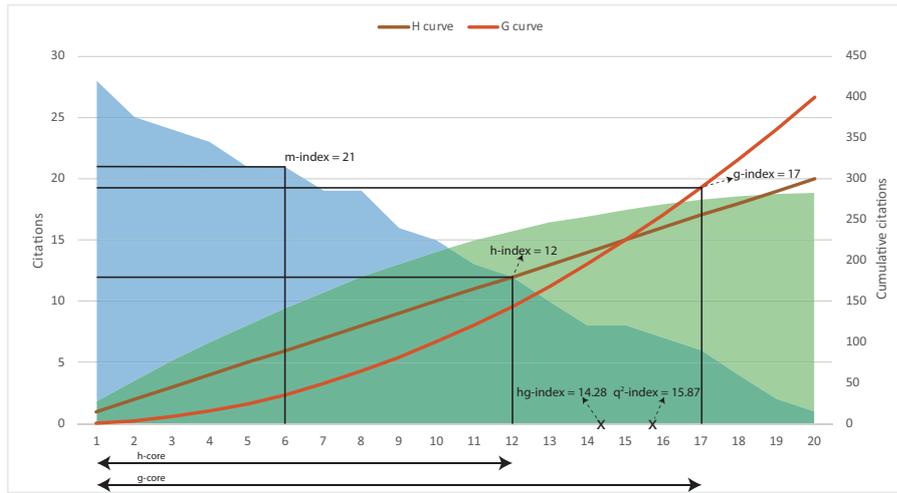
Fig. 2 Example and comparison among the four indices.

4 Citation classics

A *"citation classic"* or also called, *"classic article"* or *"literary classic"*, is a bibliometric concept introduced by Eugene Garfield [29] to designate those highly cited papers of a scientific discipline. Citation classics are as the *"gold bullion of science"* [61] and they could help us to discover potentially important information for the development of a discipline and understand the past, present and future of its scientific structure [45]. For example, it is possible to recognize the major advances in the discipline, to identify emergent or hot topics, to identify also the main intellectual markers of the research field, which could be journals or researchers or countries or research groups or institutions [29] [61]. Therefore, the development of studies on citation classics is becoming one of the most popular strategies to analyze scientific disciplines. Some examples are "Integrative & Complementary Medicine" [63], "Parkinson" [52], "Deviant Behavior" [62], "Epilepsy" [39], "Dentistry" [27] [21], "Intelligent Transportation Systems" [47], etc.

A common characteristic of studies on highly cited papers is to fix a selection criterion based on a threshold value following Garfield's recommendations [29] [31]. There exist two approaches to do it [45]: i) Setting the threshold values on the citations received [39] [52]; or ii) Setting the threshold values on the number of highly cited papers to be retrieved [27] [62] [63]. Both approaches do not take into account the citation patterns and the scientific evolution of the research areas. Therefore, the identification parameters are set according to the traditional recommendations provided by Garfield [29] [31], without considering a rigorous scientific argument and neither the circumstances of the research area when the study is done, which could introduce a bias in the choice of the highly cited papers. To overcome those problems, we introduced

in [45] the concept of H-Classics based on the popular H-index [35], which provides us an unbiased and fair criterion to build a systematic search procedure for citation classics for any field of research.

H-index was originally introduced by [35] to measure the scientific performance of a researcher through his/her publications:

> *"A scientist has index h if h of his or her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have $\leq h$ citations each."*

Burrell et al. points out that the H-index identifies the most productive core of an author's output in terms of the most cited papers [9]. For this core, consisting of the first h papers, [55] introduced the term *Hirsch core* (H-core), which can be considered as a group of high-performance publications with respect to the scientist's career [41].

Then, if we have retrieved N articles and their respective citations subject scientific category of A, we could also calculate the H-index of category A as we calculate the H-index of a researcher [45], i.e.,

> *a paper P of scientific category A is considered an H-Classic of A if and only if P is inside of the H-core of A.*

In such a way,

> *H-Classics of a research area A could be defined as the H-core of A that is composed of the H highly cited papers with more than H citations received.*

Therefore, the identification process of highly cited papers of a research area through the concept of H-Classics could be carried out in the following steps [45]:

1. *Choosing the bibliographic database to locate the scientific production and citations.* For example, Google Scholar, Scopus and WoS could be used. The latter is used in this study.
2. *Set the research area under study.* This is done by identifying those core journals that are traditionally used to disseminate scientific advances made in the area and by using two types of papers, "article" and "review". Sometimes, it is necessary to configure more complex queries in order to delimit the research area [45].
3. *Compute the H-index of the research area.* The computation of H-index of a research area is done by establishing a ranking of the papers according to their citations. If ISIWoS is used to retrieve the scientific production, it provides us filtering tools to compute easily the H-index of the research area.
4. *Compute the H-core of the research area.* This step consists in recovering the H highly cited papers that are included in the H-core of the research area, i.e., H-index= #(H-Classics). Again, we should point out that using ISIWoS this operation is easy to be carried out.

Some advantages of H-Classics to characterize the most influential papers of a research category are the following [45]:

1. It comprises in a single procedure the number of papers published in the field and the impact of those publications.
2. It provides a scientific and transparent criterion to identify the most influential papers in the scientific literature.
3. It is very simple to compute.
4. It is a criterion sensitive to the dimension of the research area.
5. And it is a criterion sensitive to the citation pattern of each research area.

5 Science Mapping Analysis Software Tools

Science mapping analysis encloses the use of different techniques and algorithms [7] to perform each one of the workflow steps. Although each step could be done separately, it would be desirable to perform the global analysis with a unique software tool. In that sense, in this section, we describe three free software tools able to perform a science mapping analysis covering the majority of the workflow stages: CitNetExplorer [24], SciMAT [17] and VOSviewer [23]. We should point out that there are other science mapping software tools [16], but not all of them cover the science mapping workflow [7][15] and are not freely available. Also, there are others bibliometric software tools but they are not focused on bibliometric network or science mapping analysis, and therefore, they are not included in our comparison. For example, Publish or Perish has not be taken into account since it is not focused on bibliometric networks.

In order to show the visualization capabilities of each software tool, an example analysis has been carried out with each software using as input the articles and reviews published by the journal *Applied Intelligence* in the period 2014-2016.

CitNetExplorer[1] (Citation Network Explorer) [24] is a software tool focused on direct citation network and based on the algorithmic historiography designed by Garfield [33]. CitNetExplorer could be used for different purposes, such as, studying the development of a research field over time, delineating research areas, studying the publication oeuvre of a researcher, or literature reviewing. Moreover, CitNetExplorer could be used to analysis patent citation network. The software tool was developed by the Centre for Science and Technology Studies at Leiden University (The Netherlands) and it is freely available.

Regarding the data management, CitNetExplorer is able to deal with large citation networks, including millions of publications and tens of millions of citation relations. Taking into account the input and output, citation networks can be imported directly into CitNetExplorer from the ISI Web of Science database, and the whole network can be exported into Pajek file format [5].

---

[1] http://www.citnetexplorer.nl/

Moreover, CitNetExplorer has a useful GUI which allows to interacting with the citation networks in order to explore it and get a deep insight of the field. The main characteristics of the GUI are:

– Visualizations of citation networks can be explored using zoom and scroll. A smart labeling algorithm ensures that labels do not overlap.
– Visualization of direct citation relations and higher-order indirect citation relations.
– Screenshots of the citation network visualization can be easily obtained and copied into a high resolution.

In addition, CitNetExplorer provides three interesting capabilities: selection of publications, drill-down and expand, and different algorithms:

– Publications can be selected simply by indicating the oldest and most recent publications of interest. Intermediate publications are automatically added to the selection.
– Drill down and expand functionality offers an intuitive way of moving through a citation network.
– Algorithm to identify connected components, clusters, core publications, and shortest and longest paths are provided.

As above mentioned, in order to illustrate the visualization capabilities of CitNetExplorer, a citation analysis was carried out using the references contained in the documents published by the journal *Applied Intelligence* during the period 2014-2016 (see Figure 3).
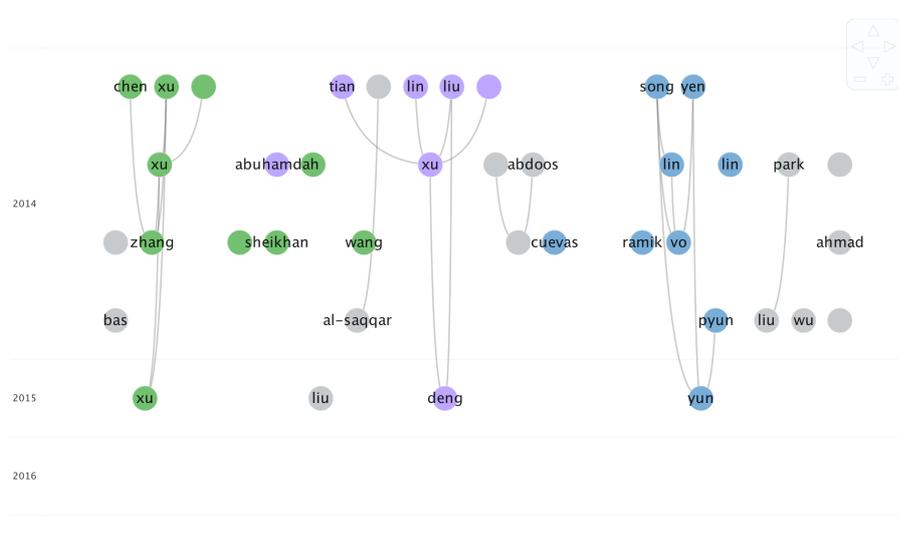
Fig. 3 CitNetExplorer Direct citation network visualization example.

SciMAT[2] (Science Mapping Analysis software Tool) [17] is an open source (GPLv3) science mapping software tool which incorporates methods, algorithms, and measures for all the steps in science mapping workflow, from preprocessing to the visualization of the results. SciMAT allows the user to carry out studies based on several kinds of bibliometric networks. Different normalization and similarity measures can be used over the data. Several clustering algorithms can be chosen to cut up the data. In the visualization module, three representations (strategic diagrams, cluster networks, and evolution areas) are jointly used, which allows the user to better understand the results. SciMAT was developed by the Secaba Lab at University of Granada (Spain) and it is freely available.

SciMAT is based on a longitudinal science mapping approach [15] which establishes the following four steps:

- To detect the substructures contained (mainly, clusters of authors, words or references) into the research field by means of a bibliometric analysis (bibliographic coupling, co-author, co-citation or co-word analysis) for each studied period.
- To lay out in a low dimensional space the results of the first step (cluster).
- To analyze the evolution of the detected clusters through the different periods studied, in order to detect the main general evolution areas of the research field, their origins, and their inter-relationships.
- To carry out a performance analysis of the different periods, clusters and evolution areas, by means of bibliometric measures.

SciMAT provides different modules that help the analyst to carry out the steps of the science mapping workflow: i) a knowledge base manager, ii) a wizard to configure and perform the analysis, and iii) a module to visualize the generated results and maps.

The main characteristics of SciMAT are:

- SciMAT incorporates all modules necessary to carry out all the steps of the science mapping workflow, which can be configured ad-hoc. It helps the analyst to carry out the different steps of the science mapping workflow, from data acquisition and preprocessing to the visualization and interpretation of the results.
- SciMAT incorporates methods to build the majority of the bibliometric networks, different similarity measures to normalize them and build the maps using clustering algorithms, and different visualization techniques useful for interpreting the output.
- SciMAT implements a wide range of preprocessing tools such as detecting duplicate and misspelled items, time slicing, data reduction and network preprocessing.
- SciMAT allows the analyst to perform a science mapping analysis in a longitudinal framework in order to analyze and track the conceptual, intellec-

---

[2] http://sci2s.ugr.es/scimat

tual or social evolution of a research field through the course of consecutive time periods.
– SciMAT builds science maps enriched with bibliometric measures based on citations such as: h-index [35], g-index [25], hg-index [3], q^2-index [10], etc.

Moreover, SciMAT incorporates a great variety of techniques and tools to perform the science mapping analysis:

– Data loaders: ISI Web of Knowledge format and RIS format.
– Bibliometric networks: co-word, author co-citation, journal co-citation, author co-citation, bibliographic coupling, journal bibliographic coupling and author bibliographic coupling.
– Preprocessing: de-duplicating (manual, by plural or by Levenshtein distance, or importing from a XML file), time-slicing, data reduction and network reduction.
– Normalization: Association Strength, Equivalence Index, Inclusion Index, Jaccards Index and Saltons Cosine.
– Mapping (through clustering): Simple Centers Algorithm, Single-linkage, Complete-linkage, Average-linkage and Sum-linkage clustering algorithms.
– Analysis: network analysis (Callon's density and centrality), performance and quality analysis (sum, minimum, maximum and average citations, and complex measures such as the h-index, g-index, hg-index or q2-index), and temporal analysis.
– Visualization: strategic diagram, cluster network, overlapping map and evolution map.
– Reports in HTML and LaTeX format.

In order to show the visualization options of SciMAT, a co-words analysis using the author keywords and ISI Keywords Plus, of the articles and reviews published by the journal *Applied Intelligence* in the period 2014-2016 has been carried out. Thus, Figure 4 shows the obtained strategic diagram.

VOSViewer[3] [23] is a software tool specifically designed for constructing and visualizing bibliometric maps, paying special attention to the graphical representation of such maps. It is appropriate to represent big maps since zoom functionality, special labeling algorithms, and density metaphors are used. The software tool was developed by the Centre for Science and Technology Studies at Leiden University (The Netherlands) and it is freely available.

VOSviewer is able to extract bibliographic network from bibliographic data. Particularly, VOSviewer accepts files in ISI Web of Science, Scopus, PubMed, and RIS format. In addition, it is able to import and export network data from GML and Pajek format. Moreover, it can generate term co-occurrence network from textual data.

Regarding the visualization, VOSviewer provides three different visualization options: network, overlay, and density. Also, it provides zoom and scroll

---

[3] http://www.vosviewver.com
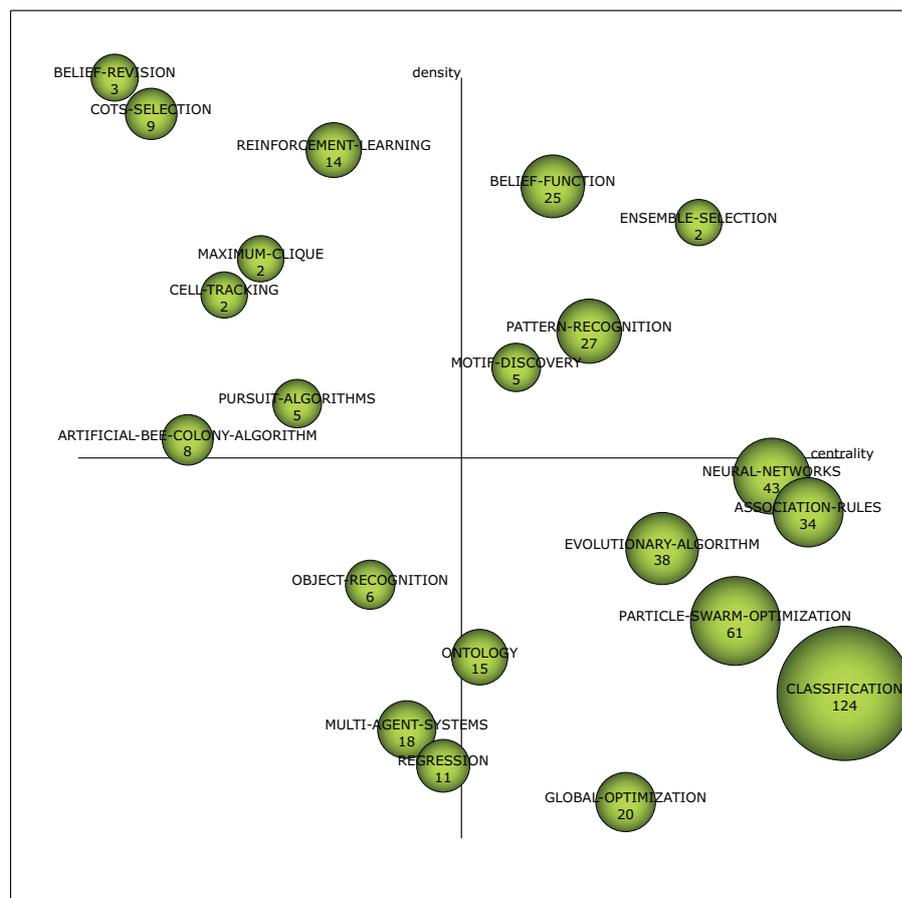
Fig. 4 SciMAT strategic diagram example

capabilities and smart labeling algorithm in order to prevent labels from over-lapping each other. Moreover, screenshots of bibliometric network visualizations can be obtained at a high resolution and can be saved in many popular graphical file formats, both bitmap and vector formats.

VOSviewer incorporates advanced layout and clustering techniques which could be fine-tuned using various parameters. Regarding the bibliometric networks, VOSViewer is able to extract a great variety, such as, co-authorship, bibliographic coupling, and co-citation networks. For instance, the influence of publications with many authors, many citations, or many references can be reduced using a fractional counting approach [51]. Moreover, data cleaning can be performed using thesaurus files. In addition, natural language processing techniques are available for creating term co-occurrence networks based on English-language textual data (e.g. relevant and non-relevant terms can be distinguished algorithmically).

Finally, a co-words analysis using the terms included in the title and abstract of the documents published by the Journal *Applied Intelligence* during the period 2014-2016 has been performed. Thus, in Figure 5 the density map obtained is shown.

Fig. 5 VOSViewer density map example

6 Conclusions

Nowadays, many institutions related to research (centers, universities, department, hospitals, etc.) need bibliometric tools to support their research policies. The publications in research are growing exponentially and we have to work to develop bibliometric tools that allow us to deal with this new big data framework that is appearing in science.

In this paper, we have presented techniques and software tools to measure the quality of the scientific production, and software tools to analyze the conceptual, social and intellectual structure of a research field. Taking into account the quality, we have presented four bibliometric indices to quantify the impact of a set of publications (e.g. author's oeuvre): h-index [35], g-index [25], hg-index [3] and q^2index [10]. In addition, it has been presented the h-classics, an approach to figure out the classic papers based on the h-index. Finally, three science mapping software tools have been introduced: CitNetExplorer [24], SciMAT [17] and VOSviewer [23].

References

1. Time aware knowledge extraction to analyze nanosafety cluster scientific activities. In: 2016 IEEE Congress on Evolutionary Computation (CEC), pp. 1233–1240. IEEE (2016). DOI 10.1109/CEC.2016.7743928

2. Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E., Herrera, F.: h-Index: A review focused in its variants, computation and standardization for different scientific fields. Journal of Informetrics **3**(4), 273–289 (2009). DOI 10.1016/j.joi.2009.04.001

3. Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E., Herrera, F.: hg-index: A new index to characterize the scientific output of researchers based on the h- and g-indices. Scientometrics **82**(2), 391–400 (2010). DOI 10.1007/s11192-009-0047-5

4. Batagelj, V., Cerinšek, M.: On bibliographic networks. Scientometrics **96**(3), 845–864 (2013). DOI 10.1007/s11192-012-0940-1

5. Batagelj, V., Mrvar: Pajek program for large network analysis. Connections pp. 47–57 (1998). DOI 10.1.1.27.9156

6. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.: Complex networks: Structure and dynamics. Physics Reports **424**(4-5), 175–308 (2006). DOI 10.1016/j.physrep.2005.10.009

7. Börner, K., Chen, C., Boyack, K.W.: Visualizing knowledge domains. Annual Review of Information Science and Technology **37**(1), 179–255 (2005). DOI 10.1002/aris.1440370106

8. Bornmann, L., Mutz, R., Daniel, H.D.: Are there better indices for evaluation purposes than theh index? A comparison of nine different variants of theh index using data from biomedicine. Journal of the American Society for Information Science and Technology **59**(5), 830–837 (2008). DOI 10.1002/asi.20806

9. Burrell, Q.L.: On the h-index, the size of the Hirsch core and Jin's A-index. Journal of Informetrics **1**(2), 170–177 (2007). DOI 10.1016/j.joi.2007.01.003

10. Cabrerizo, F., Alonso, S., Herrera-Viedma, E., Herrera, F.: q2-Index: Quantitative and qualitative evaluation based on the number and impact of papers in the Hirsch core. Journal of Informetrics **4**(1), 23–28 (2010). DOI 10.1016/j.joi.2009.06.005

11. Callon, M., Courtial, J.P., Turner, W.A., Bauin, S.: From translations to problematic networks: An introduction to co-word analysis (1983). DOI 10.1177/053901883022002003

12. Cartes-Velásquez, R., Manterola Delgado, C.: Bibliometric analysis of articles published in ISI dental journals, 2007-2011. Scientometrics **98**(3), 2223–2233 (2014). DOI 10.1007/s11192-013-1173-7

13. Cobo, M.J., Chiclana, F., Collop, A., De Oña, J., Herrera-Viedma, E.: A bibliometric analysis of the intelligent transportation systems research based on science mapping. IEEE Transactions on Intelligent Transportation Systems **15**(2), 901–908 (2014). DOI 10.1109/TITS.2013.2284756

14. Cobo, M.J., López-Herrera, A.G., Herrera, F., Herrera-Viedma, E.: A Note on the ITS topic evolution in the period 2000-2009 at T-ITS. IEEE Transactions on Intelligent Transportation Systems **13**(1), 413–420 (2012). DOI 10.1109/TITS.2011.2167968

15. Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F.: An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. Journal of Informetrics **5**(1), 146–166 (2011). DOI 10.1016/j.joi.2010.10.002

16. Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F.: Science mapping software tools: Review, analysis, and cooperative study among tools. Journal of the American Society for Information Science and Technology **62**(7), 1382–1402 (2011). DOI 10.1002/asi.21525

17. Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F.: SciMAT: A new science mapping analysis software tool. Journal of the American Society for Information Science and Technology **63**(8), 1609–1630 (2012). DOI 10.1002/asi.22688
18. Cobo, M.J., Martínez, M.A., Gutiérrez-Salcedo, M., Fujita, H., Herrera-Viedma, E.: 25 years at Knowledge-Based Systems: A bibliometric analysis. Knowledge-Based Systems **80**, 3–13 (2015). DOI 10.1016/j.knosys.2014.12.035
19. Cook, D.J., Holder, L.B.: Mining Graph Data. Wiley-Interscience (2006)
20. Costas, R., Bordons, M.: The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. Journal of Informetrics **1**(3), 193–203 (2007). DOI 10.1016/j.joi.2007.02.001
21. De la Flor-Martínez, M., Galindo-Moreno, P., Sánchez-Fernández, E., Piattelli, A., Cobo, M.J., Herrera-Viedma, E.: H-classic : a new method to identify classic articles in Implant Dentistry, Periodontics, and Oral Surgery. Clinical Oral Implants Research **27**(10), 1317–1330 (2016). DOI 10.1111/clr.12749
22. De Maio, C., Fenza, G., Loia, V., Parente, M.: Biomedical data integration and ontology-driven multi-facets visualization. In: 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2015). DOI 10.1109/IJCNN.2015.7280395
23. van Eck, N.J., Waltman, L.: Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics **84**(2), 523–538 (2010). DOI 10.1007/s11192-009-0146-3
24. van Eck, N.J., Waltman, L.: CitNetExplorer: A new software tool for analyzing and visualizing citation networks. Journal of Informetrics **8**(4), 802–823 (2014). DOI 10.1016/j.joi.2014.07.006
25. Egghe, L.: Theory and practise of the g-index. Scientometrics **69**(1), 131–152 (2006). DOI 10.1007/s11192-006-0144-7
26. Egghe, L., Rousseau, R.: An h-index weighted by citation impact. Information Processing & Management **44**(2), 770–780 (2008). DOI 10.1016/j.ipm.2007.05.003
27. Feijoo, J.F., Limeres, J., Fernández-Varela, M., Ramos, I., Diz, P.: The 100 most cited articles in dentistry. Clinical Oral Investigations **18**(3), 699–706 (2014). DOI 10.1007/s00784-013-1017-0
28. Garfield, E.: Citation Analysis as a Tool in Journal Evaluation. Science **178**(4060), 471–479 (1972). DOI 10.1126/science.178.4060.471
29. Garfield, E.: Introducing citation classics. The human side of scientific reports. Current Comments **1**(1), 5–7 (1977)
30. Garfield, E.: Citation Indexing: Its Theory and Application in Science, Technology, and Humanities. John Wiley & Sons, Inc. NY (1979)
31. Garfield, E.: 100 Citation Classics From The Journal of the American Medical Association. Journal of the American Medical Association **257**(1), 52 (1987). DOI 10.1001/jama.1987.03390010056028
32. Garfield, E.: Scientography: Mapping the tracks of science. Current Contents: Social & Behavioural Sciences **7**(45), 5–10 (1994)
33. Garfield, E., Pudovkin, A.I., Istomin, V.S.: Why do we need algorithmic historiography? Journal of the American Society for Information Science and Technology **54**(5), 400–412 (2003). DOI 10.1002/asi.10226
34. Glanzel, W.: National characteristics in international scientific co-authorship relations. Scientometrics **51**(1), 69–115 (2001). DOI 10.1023/A:1010512628145
35. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences **102**(46), 16,569–16,572 (2005). DOI 10.1073/pnas.0507655102
36. Houari, N.S., Taghezout, N.: Integrating Agents into a Collaborative Knowledge-based System for Business Rules Consistency Management. International Journal of Interactive Multimedia and Artificial Intelligence **4**(2), 61–72 (2016). DOI 10.9781/ijimai.2016.4210
37. Huang, M.H., Chang, C.P.: Detecting research fronts in OLED field using bibliographic coupling with sliding window. Scientometrics **98**(3), 1721–1744 (2014). DOI 10.1007/s11192-013-1126-1
38. Hutchins, B.I., Yuan, X., Anderson, J.M., Santangelo, G.M.: Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level. PLOS Biology **14**(9), e1002,541 (2016). DOI 10.1371/journal.pbio.1002541

39. Ibrahim, G.M., Carter Snead, O., Rutka, J.T., Lozano, A.M.: The most cited works in epilepsy: Trends in the Citation Classics. Epilepsia **53**(5), 765–770 (2012). DOI 10.1111/j.1528-1167.2012.03455.x

40. Jin, B.: h-Index: An evaluation indicator proposed by scientist. Science Focus **1**(1), 8–9 (2006)

41. Jin, B., Liang, L., Rousseau, R., Egghe, L.: The R- and AR-indices: Complementing the h-index. Chinese Science Bulletin **52**(6), 855–863 (2007). DOI 10.1007/s11434-007-0145-9

42. Kessler, M.M.: Bibliographic coupling between scientific papers. American Documentation **14**(1), 10–25 (1963). DOI 10.1002/asi.5090140103

43. Li, X., Zhou, Y., Xue, L., Huang, L.: Integrating bibliometrics and roadmapping methods: A case of dye-sensitized solar cell technology-based industry in China. Technological Forecasting and Social Change **97**, 205–222 (2015). DOI 10.1016/j.techfore.2014.05.007

44. Martínez, M.A., Cobo, M.J., Herrera, M., Herrera-Viedma, E.: Analyzing the Scientific Evolution of Social Work Using Science Mapping. Research on Social Work Practice **5**(2), 257–277 (2015). DOI 10.1177/1049731514522101

45. Martínez, M.A., Herrera, M., López-Gijón, J., Herrera-Viedma, E.: H-Classics: Characterizing the concept of citation classics through H-index. Scientometrics **98**(3), 1971–1983 (2014). DOI 10.1007/s11192-013-1155-9

46. Moed, H.F., Bruin, R.E., Leeuwen, T.N.: New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. Scientometrics **33**(3), 381–422 (1995). DOI 10.1007/BF02017338

47. Moral-Muñoz, J.A., Cobo, M.J., Chiclana, F., Collop, A., Herrera-Viedma, E.: Analyzing Highly Cited Papers in Intelligent Transportation Systems. IEEE Transactions on Intelligent Transportation Systems **17**(4), 993–1001 (2016). DOI 10.1109/TITS.2015.2494533

48. Moral-Muñoz, J.A., Cobo, M.J., Peis, E., Arroyo-Morales, M., Herrera-Viedma, E.: Analyzing the research in Integrative Complementary Medicine by means of science mapping. Complementary Therapies in Medicine **22**(2), 409–418 (2014). DOI 10.1016/j.ctim.2014.02.003

49. Murgado-Armenteros, E.M., Gutiérrez-Salcedo, M., Torres-Ruiz, F.J., Cobo, M.J.: Analysing the conceptual evolution of qualitative marketing research through science mapping analysis. Scientometrics **102**(1), 519–557 (2015). DOI 10.1007/s11192-014-1443-z

50. Noyons, E.C.M., Moed, H.F., Luwel, M.: Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. Journal of the American Society for Information Science **50**(2), 115–131 (1999). DOI 10.1002/(SICI)1097-4571(1999)50:2¡115::AID-ASI3¿3.0.CO;2-J

51. Perianes-Rodriguez, A., Waltman, L., van Eck, N.J.: Constructing bibliometric networks: A comparison between full and fractional counting. Journal of Informetrics **10**(4), 1178–1195 (2016). DOI 10.1016/j.joi.2016.10.006

52. Ponce, F.A., Lozano, A.M.: The most cited works in Parkinson's disease. Movement Disorders **26**(3), 380–390 (2011). DOI 10.1002/mds.23445

53. van Raan, A.F.J.: Measuring Science. In: H.F. Moed, W. Glänzel, U. Schmoch (eds.) Handbook of Quantitative Science and Technology Research, chap. Measuring, pp. 19–50. Springer Netherlands (2005)

54. Rodriguez-Ledesma, A., Cobo, M.J., Lopez-Pujalte, C., Herrera-Viedma, E.: An overview of animal science research 1945-2011 through science mapping analysis. Journal of Animal Breeding and Genetics **132**(6), 475–497 (2015). DOI 10.1111/jbg.12124

55. Rousseau, R.: New developments related to the Hirsch index. Science Focus **1**(4), 23–25 (2006)

56. SCImago: SJR - SCImago Journal & Country Rank (2007)

57. Settouti, N., Bechar, M.E.A., Chikh, M.A.: Statistical Comparisons of the Top 10 Algorithms in Data Mining for Classi cation Task. International Journal of Interactive Multimedia and Artificial Intelligence **4**(1), 46–51 (2016). DOI 10.9781/ijimai.2016.419

58. Sidiropoulos, A., Katsaros, D., Manolopoulos, Y.: Generalized Hirsch h-index for disclosing latent facts in citation networks. Scientometrics **72**(2), 253–280 (2007). DOI 10.1007/s11192-007-1722-z

59. Small, H.: Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science **24**(4), 265–269 (1973). DOI 10.1002/asi.4630240406
60. Small, H.: Visualizing science by citation mapping. Journal of the American Society for Information Science **50**(9), 799–813 (1999). DOI 10.1002/(SICI)1097-4571(1999)50:9¡799::AID-ASI9¿3.0.CO;2-G
61. Smith, D.: Ten citation classics from the new zealand medical journal. New Zealand Medical Journal **120**(1267), 2871–2875 (2007)
62. Stack, S.: Citation classics in deviant behavior: A research note. Deviant Behavior **34**(2), 85–96 (2013). DOI 10.1080/01639625.2012.707539
63. Tam, W.W., Wong, E.L., Wong, F.C., Cheung, A.W.: Citation classics in the integrative and complementary medicine literature: 50 frequently cited articles. European Journal of Integrative Medicine **4**(1), e77–e83 (2012). DOI 10.1016/j.eujim.2011.12.004
64. Tang, L., Shapira, P.: China-US scientific collaboration in nanotechnology: patterns and dynamics. Scientometrics **88**(1), 1–16 (2011). DOI 10.1007/s11192-011-0376-z
65. Van Eck, N.J., Waltman, L.: How to normalize cooccurrence data? An analysis of some well-known similarity measures. Journal of the American Society for Information Science and Technology **60**(8), 1635–1651 (2009). DOI 10.1002/asi.21075
66. Vanclay, J.K.: On the robustness of theh-index. Journal of the American Society for Information Science and Technology **58**(10), 1547–1550 (2007). DOI 10.1002/asi.20616