International Journal of Technology (2015) 4: 650-658
ISSN 2086-9614

BIG DATA ANALYSIS OF INDONESIAN SCHOLARS' PUBLICATIONS: A RESEARCH THEME MAPPING

Isti Surjandari[1*], Arian Dhini[1], Esther Widya Impola Lumbantobing[1], Anita Titiani Widari[1], Irfan Prawiradinata[2]

[1] *Department of Industrial Engineering, Faculty of Engineering, Universitas Indonesia, Kampus Baru UI Depok, Depok 16424, Indonesia*
[2] *Department of Economics, Faculty of Economics and Business, Universitas Indonesia, Kampus Baru UI Depok, Depok 16424, Indonesia*

(Received: September 2015 / Revised: October 2015 / Accepted: October 2015)

ABSTRACT

The advancement of science and technology plays an important role in enhancing the international competitiveness of a nation. One of the indicators for measuring the advancement in science and technology of a nation is the amount of research that is published in international journals. In order to improve the quantity and quality of research publications, research themes mapping of published studies is required. By conducting research themes mapping, the research road map for future research can be developed effectively. In this study, Indonesian scholars' publications, which are indexed in Scopus during the last five years (2010–2014), were analyzed by applying co-word analysis. Co-word analysis, which is part of text mining, is applied to find the interrelationship among research themes in a big corpus of data. As a result, there are fifteen main research themes in Indonesia as on Quadrant I and IV of strategic diagram.

Keywords: Bibliometrics; Big data analysis; Co-word analysis; Indonesia; Research theme mapping

1. INTRODUCTION

Science and technology are two powerful drivers in responding to the problems of developing countries. Scientific research plays an important role in creating economic growth (Nguyen & Pham, 2011), while technological innovation has also become the main source of increased productivity, the major tool of economic competition in the world market (Wang et al., 2007; Apak & Atay, 2015). In other words, the advancement of science and technology is a key driver of a nation to survive in global competition.

In relation with global competitiveness, Indonesia is still lagging behind some ASEAN countries; with Singapore, Malaysia, and Thailand taking the lead over Indonesia (Schwab, 2014). The number of Indonesian research publications, as the output of scientific research, is also falling behind those countries. The highest position in terms of total research publications was occupied by Singapore, followed by Malaysia in second place, Thailand in third place and Indonesia in fourth place (SCImago, 2015). Therefore, based on these facts, the Indonesian government is required to establish appropriate policies in science and technology, in order to enhance the global competitiveness of Indonesia.

* Corresponding author's email: isti@ie.ui.ac.id, Tel. +62-21-78888805, Fax. +62-21-78885656
Permalink/DOI: http://dx.doi.org/10.14716/ijtech.v6i4.1956

The Indonesian government has developed a National Strategic Policies in Science & Technology (Jakstranas IPTEK) for the period of 2010–2014. One of government policies related to Jakstranas IPTEK is the arrangement of seven (7) focus areas for the development of science and technology in Indonesia, i.e. (i) food security; (ii) energy; (iii) information and communication technology; (iv) technology and transportation management; (iv) defense and security technology; (vi) health technology and medicine; and (vii) advanced materials. Through this arrangement, one of the objectives to be achieved by the government is to increase the number of international publications from Indonesia (LIPI, 2011). Although the seven research focus areas have been determined, it was still not clear whether those seven themes represent the overall scope and breadth of Indonesian research publications. Thus, a research mapping exercise is necessarily required to determine the main research themes from Indonesian research publications.

Bibliometric analysis has been widely applied in order to gain knowledge of research domains, trends and structures (Oh & Lee, 2014). Bibliometrics covers measurement of document properties and document-related processes (Borgman & Furner, 2002). Bibliometrics has overlapping with scientometrics, the science measurement field, in terms of its applications on science-related documents (Thelwall, 2008). Traditionally, bibliometric analysis was conducted through the narrow body of literature selection or journal selections from a narrow topic, which produced less accurate results (Song & Kim, 2013). However, in this 'Big Data' era, where scientific databases and computer power has been evolving and growing dramatically in terms of volume and velocity, big data analysis can be applied to overcome the previous approach limitation.

Text mining, a mining technique from unstructured data, is a technique in big data analysis which can be applied for bibliometric analysis in a more comprehensive way to overcome the limitation of the traditional approach. In order to identify the research theme mapping of Indonesian scholars, text mining was selected. In this research, co-word analysis was applied as a text mining technique to seek valuable information from the big corpus of data in research publications, and transform it into valuable knowledge.

2. METHODOLOGY

Content analysis, which is part of text mining, is a systematic and replicative method to reduce the words of a text into several categories of content based on specific coding rules (Berelson, 1952; Krippendorff, 1989). It was often applied to analyze changes in trends related to the content of scientific theories and methodological approaches, by analyzing the content of an article in a journal within the theme of a particular discipline (Loy, 1979).

Co-word analysis is one of content analysis techniques that offer a significant methodological approach to knowledge discovery (Law et al., 1986). Co-word analysis maps the literature based on the interaction of keywords that are the representation of the important concept of a research article (Delecroix & Epstein, 2004; Wu & Leu, 2014). Co-word analysis uses the pattern of the words or noun phrases co-occurrence to identify relationships among ideas in a field of study.

The data used in this study is a collection of 5,878 research articles published by eight of the best universities in Indonesia, based on 2015 QS World University Ranking and government research institutes conducted under the Ministry of Research and Higher Education, which has been indexed in Scopus. Scopus is one of the largest multidisciplinary databases of scientific literature in the world (Bar-Ilan, 2008). The articles were published in the last 5 years, from 2010 until 2014.

Data pre-processing was conducted in the following steps: (1) adding the keyword and (2) the word grouping. The keyword is a very important component in co-word analysis. If there was an article which had no keywords or the keywords were not representative of the article, then the keywords were added manually. After all keywords were collected, a grouping based on keyword similarities was performed. In this study, each keyword was grouped by a subject category or a branch of science as stated on SCImago Journal and Country Rank. For example, keywords such as '*Drug', 'Drug-effect', 'Drug-efficacy',* and other words or phrases that are related to drugs are grouped into the *Pharmacology* word group.

For the purpose of this study, SciMAT, a science mapping analysis software tool, was used to perform the co-word analysis. Co-word analysis is able to find the interrelationship among research themes based on the co-occurrence of keywords. A cluster network, which describes the relationship among keywords in a research theme, can be created through development of the co-occurrence matrix. The matrix was then normalized and the similarity was measured using an index. The index used in the calculation of a co-occurrence matrix similarity is an equivalence index.

An Equivalence Index (E_{ij}) measures the probability of keyword '*i'* appearing simultaneously in a document where keyword '*j'* existed and vice versa (He, 1999) which is shown in Equation 1.

$$E_{ij} = (C_{ij} / C_i) \times (C_{ij} / C_j) = (C_{ij})^2 / (C_i \times C_j) \quad (1)$$

where C_{ij} is the number of documents in which there is a keyword pair (*i* and *j*), C_i is the keyword frequency of *i* on all documents, C_j is the keyword frequency of *j* on all documents.

Keywords which have a high equivalence index were chosen as the central theme of a cluster network. The depth of a cluster network to be formed is then specified using a simple center algorithm with a threshold value (Cobo et al., 2012). The threshold value is the minimum and maximum number of members in a cluster, which is determined subjectively based on the desired depth of analysis (López-Herrera et al., 2010).

After all clusters are developed, each central theme is visualized on a strategic diagram. Strategic diagram is used to illustrate the "local" and "global" context of research themes (Law et al., 1988). The *x*-axis shows the strength of the global context, while the *y*-axis shows the strength of the local context. Density and centrality are the parameters used to measure the local and global strength respectively. Those two parameters are described on Equations 2 and 3,

$$d = 100(\sum {}^{e_{ij}}/_{x}) \quad (2)$$

$$c = 10 * \sum e_{kh} \quad (3)$$

where *i* and *j* are keywords belonging to the theme, *x* is the number of keywords in the theme, *k* is a keyword belonging to the theme and *h* is a keyword belonging to other themes.

$$p = p_m + p' \quad (4)$$

$$\rho = \rho_m + \rho' \quad (5)$$

$$T = T_m + T' \quad (6)$$

3. RESULTS AND DISCUSSION

The results of this study are shown in the visualisation of a strategic diagram in Figure 1. From the strategic diagram, it can be seen that over the last five years there were 28 clusters of research themes which have been the most highlighted by academicians and researchers in Indonesia. In order to measure the performance or productivity of the identified themes, a quantitative measure of the number of published documents associated with each research theme can be seen from the sphere size and label in Figure 1. Based on the number of publications, the five most productive research themes are pharmacology, organic chemistry, children welfare, virology and hematology.

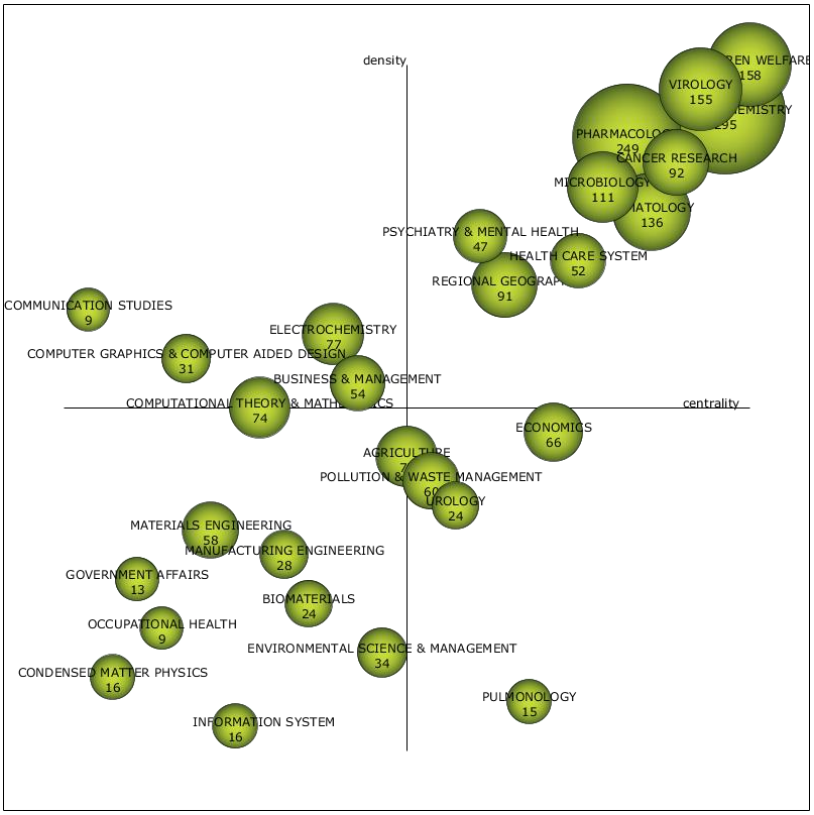

Figure 1 Strategic diagram for the 2010-2014 period

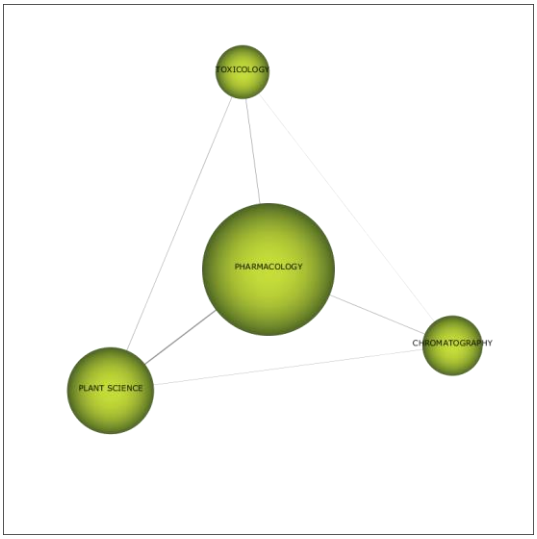

Figure 2 Cluster Network of Pharmacology

Each identified theme in the strategic diagram is actually a cluster network of subthemes. A research theme is explained by a central theme and supported by themes surrounding the central theme (sub-themes). An example of a cluster network of Pharmacology theme is shown in Figure 2. Pharmacology acts as the central theme and it is related to sub-themes such toxicology, plant science, and chromatography. The thickness of the link between the spheres is proportional to the equivalence index Eij. The central theme with its interrelated sub-themes is important for defining a distinct research theme. Further definition of all formed research themes is explained in the next sub section.

Each of research themes is discussed thoroughly in this section. Based on the strategic diagram in Figure 1, four kinds of themes were found in each quadrant. The themes in Quadrant I and IV are considered as the main themes because they are important for structuring other research themes (Cobo et al., 2015).

3.1. Main Themes of Quadrant I

Themes in Quadrant I are both well developed and important for the structuring of other research areas. They are known as the motor-themes of the specialty, given that they present strong centrality and high density. Figure 1 shows that there were ten clusters of motor themes found in this study. These motor themes were defined by considering the central themes and sub-themes that formed each cluster network as shown in Table 1. Each cluster is labelled and explained as follows:

Cluster I: Child health and well-being

The central theme of this cluster is *child welfare* with keywords such as child development, child growth, and child protection. Other mediator themes that support this central theme are *pediatrics* (e.g., child health, childhood leukemia, childhood disease); *education and curriculum* (e.g., preschool child, e-learning, curriculum); and *demography* (e.g., adolescent, age distribution, population size). Given that its topics are related not only to child welfare, but also to children health, education, and demography, this cluster is defined as child health and well-being.

Cluster II: Pharmacology

The central theme of this cluster is a group of keywords in *pharmacology*, e.g. drug effect, drug efficacy, drug safety. The central theme is supported by sub-themes: *toxicology* (e.g., drug cytotoxicity, toxicity testing, cytotoxic); *botany* (e.g. medicinal plant, plant extract, plant root); *chromatography* (e.g. column chromatography, liquid chromatography, gas chromatography). Aspects of pharmacology are discussed as focus topics in the literature of this cluster.

Cluster III: Genome

The theme of this cluster is revolving around the genome, the genetic material of an organism. The central theme is *organic chemistry* with high frequency keywords such as protein, protein expression, protein binding. The central theme is closely related to sub-themes: *genetics* (e.g. nucleotide sequence, gene expression, genetic variability); *cell biology* (e.g. cell structure, cell proliferation, cell nucleus); and *biochemistry* (e.g. structural biochemistry, enzyme inhibition, nucleic acid).

Cluster IV: Infectious disease

This cluster is formed by the central theme *virology* (e.g. dengue virus, virus transmission, virus isolation). It is strongly related to sub-theme of *infectious disease* (e.g. HIV infection, Hepatitis, malaria). Other related virology sub-themes are *immunology* (e.g. viral antibody, immunohistochemistry, immune response); and *epidemiology* (e.g. epidemic, pandemic, disease progression). Given that the topics in this cluster are related not only to virology, this cluster is defined as infectious disease, as it defines a broader and more distinct research area.

Cluster V: Psychology and mental health

Psychiatry is the central theme of this cluster with keywords e.g. mental health, schizophrenia, depression. It is closely related to the sub-theme *social psychology* (e.g. perception, social status, attitude). Because these two disciplines are closely related, the theme of this cluster is defined as psychology and mental health (psychiatry). The mediator themes *neurology* (e.g. neurologic disease, short term memory, cognition) and *sociology* (e.g. social behavior, social development, social structure) support the theme of psychology and psychiatry.

Cluster VI: Internal medicine

The internal medicine theme is defined by the central theme of *hematology* (e.g. blood, blood glucose, hemoglobin blood level) which is closely related to mediator themes of *cardiology* (e.g. cardiovascular disease, coronary disease, heart infarction), *endocrinology* (e.g. diabetes, insulin, diabetes mellitus), and *hepatology* (e.g. liver, liver cirrhosis, liver toxicity). This theme is related to different aspects of the medical specialty dealing with adult diseases.

Cluster VII: Health care system and public health

The central of this cluster is *health care system* (e.g. health care quality, health care policy, health service). Other mediator themes that support this central theme are *public health* (e.g. attitude to health, delivery of health care, public health service); *nursing care* (nurse, nurse attitude, nurse's role); and *intensive & critical care* (e.g. intensive care unit, emergency care, emergency medicine). Given that its topics are related not only to health care system, but also to public health, this cluster is defined as a health care system and public health.

Cluster VIII: Cancer research

The cancer research theme (e.g. cancer staging, cancer cell culture, cancer combination chemoteraphy) represents research-related to cancer causes and treatments. The central theme *cancer research* is supported by the theme *oncology* (e.g. tumors, tumor marker, oncologic surgery), *molecular biology* (e.g. DNA sequence analysis, mutation, neoplasm) and *therapy* (e.g. monotherapy, chemotherapy, chemo radiotherapy).

Cluster IX: Microbiology

This cluster is developed by the central theme of *microbiology* (e.g. microbiology, microorganisms, microbial activity), which deals with microscopic organisms. Three mediator themes which support this cluster are *bacteriology* (e.g. bacteria, escherichia coli, bacterial strain), *mycology* (e.g. antifungal, fungal extract, basidiomycota), and *physiology* (e.g. physical activity, microbial physiology, pathophysiology). All of them are classified as Microbiology.

Cluster X: Urban and regional studies

The central theme of this cluster is regional geography (e.g. Indonesia, Sumatra, East Java). It is described further by mediator themes of *spatial analysis* (e.g. remote sensing, GIS, satellite imagery) and *forestry* (e.g. forestry, agroforestry, forest management). *Urban studies* (e.g. urban planning, mass transportation, urbanization) is a unique mediator theme that is closely related to the central theme. Given that its focal topics are related not only to regional studies, but also to urban studies, this cluster is defined as urban and regional studies.

3.2. Main Themes of Quadrant IV

Other than motor themes in Quadrant I, themes in Quadrant IV are also considered as the main themes. Research themes in Quadrant IV basically have great influence on other research themes, although they are not yet well developed. Five clusters of themes which belong to this quadrant are known as basic and transversal themes and are further labelled by considering the sub-themes as shown in Table 2.

Cluster I: Agriculture

This cluster is formed by the central theme of *agriculture* (e.g. crops, poultry, cattle) which discusses the cultivation of animals, plants, and other life forms for food, biofuel, medicinal and

other products. Three mediator themes which support this cluster are *food technology* (e.g. vegetable oil, virgin coconut oil, food analysis), *animal science* (e.g. animal disease, animal food, animal product), and *renewable energy* (e.g. biodiesel, biogas, crude palm oil) which are topics in agriculture technology and management.

Cluster II: Economics and international relations

The central theme *economics* (e.g. economic development, socioeconomic, global economy) is strongly related to the sub-theme of *international relations* (e.g. international cooperation, international trade, ASEAN community). Other sub-themes which are related to the central theme are *policy studies* (e.g. foreign policy, trade policy, policy making) and cultural studies (e.g. cultural factor, cross cultural, ethnic difference). In general, the research theme studies two distinct topics : economics and international relations.

Cluster III: Pollution and waste management

This cluster is formed by the central theme *pollution and waste management* which is explained by keywords such as waste water, air pollution, recycling, etc. Three mediator themes which support this cluster are *water science and technology* (e.g. water quality, groundwater, water treatment), *fuel technology* (e.g. biofuel, fuel cell, alternative fuel), and *biotechnology* (e.g. bioremediation, biosynthesis, bioequivalence). All mediator themes of this cluster support the central theme, which explains the technology in pollution and waste management.

Cluster IV: Obstetrics and genitourinary medicine

This cluster is labelled as obstetrics and genitourinary medicine because it is formed by the central theme *urology* (e.g. diuretic, testosterone, and urine) and it is strongly related to a distinct sub-theme *obstetrics* (e.g. pregnancy, preeclampsia, and uterus). The sub-themes *gynecology* (e.g. vagina, estrogen, ovary cyst) and *nephrology* (e.g. kidney, renal dialysis, kidney failure) formed the genitourinary medicine.

Cluster V: Respiratory Medicine

Research in this cluster specializes in aspects related with pulmonary and respiratory medicine. It covers wide topics such as *pulmonology* (e.g. lung tuberculosis, pneumonia, pulmonary hypertension), *otorhinolaryngology* (e.g. auditory simulation, hearing impairment, sinusitis), and *biomedical engineering* (e.g. artificial respiration, biomedical implants, biomedical materials) which seek to understand various diseases and treatments on respiratory system.

3.3. Main Themes of Quadrant II

Themes in Quadrant II have well-developed internal relationships (high density), but are considered of marginal importance to other research areas (low centrality). Based on the strategic diagram in Figure 1, four research themes were categorized as specialized and peripheral themes which were labelled as follows:

1) Communication studies
2) Electrical engineering
3) Business and management
4) Computer graphics and aided design

3.4. Main Themes of Quadrant III

Themes in Quadrant III are both weakly developed and marginal. The themes of this quadrant have low density and low centrality, mainly representing either emerging or disappearing themes. There were nine clusters of research themes which belong to Quadrant III:

1) Metallurgy and materials
2) Logistics and Manufacturing
3) Politics and Governance
4) Nuclear Physics

5) Biomaterials
6) Environmental science and management
7) Information system
8) Occupational Health
9) Computational Intelligence

By applying big data analysis in this study, a comprehensive result of Indonesian scholars' research themes was identified. From the large number of publications in the last five years, Indonesian scholars' covered twenty-eight research themes, in which fifteen of them were the main research themes, i.e. Child welfare, pharmacology, organic chemistry, virology, psychiatry, hematology, health care systems, cancer research, microbiology, regional geography, agriculture, economics, population and waste management, urology and pulmonology, as shown in Quadrants I and IV. From these fifteen main research themes, it can be concluded that researches in Indonesia during the last five years were highly focused on health care & medicine. Most of the themes in those quadrants were matched to four of seven research focus areas in Indonesia as issued by the government, namely the fields of health technology and medicine, food security, transportation management and energy. However, two research focus areas, such as the fields of information technology and communication and advanced materials, were not considered as main themes. They were discussed in Quadrants II and III.

Hence, these results give a clear description of Indonesian scholars' research strengths as well as weaknesses. This research theme mapping can be used as input for evaluating the existing seven research focus areas in order to develop an effective research road map

4. CONCLUSION

By mapping research publications using co-word analysis, it was found that there were twenty-eight research themes, which have been highlighted by academicians and researchers in Indonesia during the 2010–2014 period. From these themes, fifteen were selected as the main research themes in Indonesia, which cover a broad spectrum of research areas. Most of these themes have been set by the government as focus areas for Indonesian research. Moreover, this study also shows that more specialized and distinct themes occurred as main research areas, such as child well-being, genome, microbiology, urban and regional studies, economics and international relations, and waste management. These findings can be used as a base for an evaluation of existing research focus areas in Indonesia and for the road map development of future research.

5. ACKNOWLEDGEMENT

The authors would like to express their gratitude and appreciation to Universitas Indonesia for financing this study under the Multidisciplinary Research Grant, No 1651/UN2.R12/HKP.05.00/2015.

6. REFERENCES

Apak, S., Atay, E., 2015. Global Competitiveness in the EU through Green Innovation Technologies and Knowledge Production. *Procedia-Social and Behavioral Sciences,* Volume 181, pp. 207–217

Bar-Ilan, J., 2008. Which h-index?—A comparison of WoS, Scopus and Google Scholar. *Scientometrics,* Volume 74(2), pp. 257–271

Berelson, B., 1952. *Content Analysis in Communication Research*. New York, US, The Free Press

Borgman, C.L., Furner, J., 2002. Scholarly Communication and Bibliometrics. *Annual Review of Information Science and Technology,* Volume 36, pp. 3–72

Cobo, M., Martínez, M., Gutiérrez-Salcedo, M., Fujita, H., Herrera-Viedma, E., 2015. 25years at Knowledge-based Systems: A Bibliometric Analysis. *Knowledge-Based Systems,* Volume 80, pp. 3–13

Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F., 2012. SciMAT: A New Science Mapping Analysis Software Tool. *Journal of the American Society for Information Science and Technology,* Volume 63(8), pp. 1609–1630

Delecroix, B., Epstein, R., 2004. Co-word Analysis for the non-Scientific Information Example of Reuters Business Briefings. *Data Science Journal,* Volume 3, pp. 80–87

Krippendorff, K., 1989. *Content Analysis: An Introduction to its Methodology*. Newbury Park, Sage Publications.

Law, J., Bauin, S., Courtial, J., Whittaker, J., 1988. Policy and the Mapping of Scientific Change: A Co-word Analysis of Research into Environmental Acidification. *Scientometrics,* Volume 14(3–4), pp. 251–264

Law, J., Rip, A., Callon, M., 1986. *Mapping the Dynamics of Science and Technology: Sociology of science in the Real World*, Macmillan

LIPI, 2011. *Indikator IPTEK Indonesia 2011*. Jakarta, Pusat Penelitian dan Pengembangan IPTEK (Pappiptek)

López-Herrera, A.G., Cobo, M.J., Herrera-Viedma, E., Herrera, F., 2010. A Bibliometric Study about the Research based on Hybridating the Fuzzy Logic Field and the Other Computational Intelligent Techniques: A Visual Approach. *International Journal of Hybrid Intelligent Systems,* Volume 7(1), pp. 17–32

Loy, P., 1979. Content Analysis of Journal Articles as a Technique for Historical Research. *Journal of the History of Sociology,* Volume 1(2), pp. 93–101

Nguyen, T.V., Pham, L.T., 2011. Scientific Output and its Relationship to Knowledge Economy: An Analysis of ASEAN Countries. *Scientometrics,* Volume 89(1), pp. 107–117

Oh, J., Lee, B.G., 2014. A Technical Approach for Suggesting Research Directions in Telecommunications Policy. *KSII Transactions on Internet and Information Systems (TIIS),* Volume 8(12), pp. 4467–4488

Schwab, K., 2014. *The Global Competitiveness* Report 2014–15. Switzerland: World Economic Forum

SCImago, 2015. *SJR - SCImago Journal and Country Rank.* Available online at http://www.scimagojr.com, Accessed on 29/05/2015

Song, M., Kim, S., 2013. Detecting the Knowledge Structure of Bioinformatics by Mining Full-text Collections. *Scientometrics,* Volume 96(1), pp. 183–201

Thelwall, M., 2008. Bibliometrics to Webometrics. *Journal of Information Science,* Volume 34(4), pp. 605–621

Wang, T.-Y., Chien, S.-C., Kao, C., 2007. The Role of Technology Development in National Competitiveness—Evidence from Southeast Asian countries. *Technological Forecasting and Social Change,* Volume 74(8), pp. 1357–1373

Wu, C.-C., Leu, H.-J., 2014. Examining the Trends of Technological Development in Hydrogen Energy using Patent Co-word Map Analysis. *International Journal of Hydrogen Energy,* Volume 39(33), pp. 19262–19269