Using Bibliometric Analysis and Text Mining to Improve the Thai Talent Database

Alisa Kongthon, Choochart Haruechaiyasak, Kanokorn Trakultaweekoon
National Electronics and Computer Technology Center (NECTEC)
National Science and Technology Development Agency, Thailand Science Park, Pathumthani, Thailand

Abstract--**In today's competitive world, "Talentism has become the new capitalism." Knowledge workers are viewed as valuable assets to their organizations and nations. In developing countries such as Thailand in particular, the number of experts in science and technology is quite limited. The mobility of talent between academia, government and industry is therefore essential for knowledge transfer and technology diffusion. In recognition of this fact, the Royal Thai Government developed a National Talent Data Base ("Talent Database") in 2014, in order to facilitate mobilization of necessary talent and skills. However, the current Talent Database only supports keyword searches for potential candidates. Keyword searches alone provide no indication of the associations between researchers that may better allow talent managers to better pinpoint necessary and related skill sets, and to locate all potential candidates for positions. In this paper, we suggest improvements that may make the Talent Database more useful to managers seeking to plan talent mobility for R&D activities in Thailand. As an illustration, we apply "bibliometric analysis" and a "text mining" approach to identify topical emphases and experts' knowledge networks in the field of data science. The results could be used to assist decision makers to better match the demand with the right talents.**

I. INTRODUCTION

"The success of any national or business model for competitiveness in the future will be placed less on capital and much more on talent. We could say that the world is moving from capitalism to talentism [36]," says Klaus Schwab, Founder and Executive Chairman of World Economic Forum. And yet again in 2015, he stated that talent, not capital, will be the key factor linking innovation, competitiveness and growth in the 21st century [37]. This reflects a global paradigm shift from "capitalism" to "talentism." Despite today's high unemployment rates, there is a shortage of skilled workers on a global scale [9]. As a result, talent mobility has become an essential part of strategic talent management. According to the Organization for Economic Co-operation and Development (OECD), talent mobility is essential for the economic growth of nations because the movement of skilled workers contributes to the creation and diffusion of both codified and tacit knowledge [23]. Codified knowledge can be easily transferred as information through documentation, academic papers, or technical notes. Tacit knowledge, on the other hand, can be effectively transferred among individuals with a common social interaction and physical proximity. Hence, mobilized scientists or researchers can transfer both their codified and tacit knowledge to their colleagues and those in close contact. The topic of talent mobility, has gained a remarkable degree of interest in academia, government and industry over the past few years. Several works have focused on national or regional talent mobility in various countries. For instance, the work of Reiner conceptualized a framework on talent mobility from a European perspective [33]. Oishi addressed the immigration policies on highly skilled migrants in Japan [24, 25]. The work of Puteh, Nor, and Zulkifli focused on the mobility of young talent in Malaysia [32].

In this paper, we focus on a talent mobility initiative in Thailand. Thailand is a developing country with a limited number of experts in science and technology fields. More importantly, 83% of talented and skilled manpower in Thailand work for government agencies and universities, while only 17% work for the private sector [20]. A survey conducted by the National Science Technology and Innovation Policy Office has also found that few private companies have collaborations with government agencies and universities [21]. For this reason, in order to strengthen Thailand's competitiveness, the Talent Mobility[1] program was initiated in 2014. The program finally got approval from the Cabinet in 2015 [20]. The Talent Mobility program aims to enable mobility of researchers at government agencies and universities, to assist the private sector in technological upgrading and innovation. The researchers are authorized to work full-time or part-time with industry for up to 2 years.

The first task in the Talent Mobility program was to create a national Talent Database. Talent mobility manager would use this database to identify experts in a specific area requested by industry. The Talent Database was initially constructed using the Researcher Profile Database from the Thai National Research Repository ("TNRR"). TNRR is an initiative to provide access to research output created by researchers from government agencies, universities, and research institutes [35]. The database was constructed based upon voluntary responses from researchers to create a profile listing first name, last name, education, work history, and area of expertise. There were 17,863 researchers in the initial list. After exploring this database, we found that the area of expertise field is empty for the majority of entries. The area of expertise is a very important field in the Talent Database because it is the primary field to use in the matching the talent demand and supply. In fact, area of expertise was listed for only 5644 researchers in the database.

Currently, when talent mobility manager wants to search for an expert from the Talent Database, she has to perform a keyword search in a particular field of expertise. The drawback of a keyword search is that one can only find

[1] http://talentmobility.or.th

records containing the query terms. For instance, if the user wants to search for experts in “data mining”, the current database would return records where the field of expertise contains the term “data mining” but not records with related terms such as “machine learning” or “pattern recognition”. Moreover, with the keyword search, only the list of experts will be retrieved. The associations between co-authorship patterns cannot be detected. As a result, we proposed to apply bibliometric analysis and text mining to help identify experts and their networks in order to provide a better match between talent demand and the available supply of experts.

In the remainder of this paper, Section II presents the related work in the area of bibliometric analysis, and text mining. We then illustrate the applications of our proposed method through a case study identifying experts in the area of data science in Section III. The discussion of the case is presented in Section IV. The conclusions are discussed in the last section.

II. RELATED WORK

Bibliometrics is a study that uses statistical and mathematical methods to analyze literature patterns [18]. The pioneers in bibliometrics include Derek de Solla Price [31], Eugene Garfield [5] and Henry Small [34]. The commonly used bibliometric methods are content analysis and citation analysis. Content analysis uses the co-occurrence of keywords or terms in the documents on a given subject to discern relationships among them [2]. If certain terms tend to appear together in documents, this is taken as evidence of a possible relationship. While content analysis provides an immediate picture of actual research content within the literature, citation analysis, provides information on the direction and flow of research. Small introduced the concept of co-citation analysis and defined it as “the frequency with which two items of earlier literature are cited together by the later literature. [34]” Co-citation analysis has been successfully applied to identify cognitive structure of many disciplines. Cognitive structure provides information on the direction and flow of scientific thought.

With advancements in information retrieval, computational linguistics, natural language processing, and knowledge discovery in databases, bibliometric analysis has transitioned into text mining where the abstract and full text of a document maybe explored [10]. Such analytical approaches can be used to infer knowledge from a body of literature and provide insights for researchers and practitioners. For instance, these approaches can assist in mapping and profiling research domains [1, 16, 26, 27, 29] as well as exploring a research community and networks [6, 7, 8, 30]. Also recently, Kostoff has introduced the concept of Literature-Related Discovery and Innovation (LRDI) which is a text mining approach for bridging unconnected disciplines to hypothesize radical discovery. [11, 12]. LRDI focuses mainly on the medical domain such as Raynaud’s Phenomenon [13], multiple sclerosis [14], and chronic kidney disease [15].

III. A CASE STUDY PROFILING TALENT SUPPLY IN THE AREA OF DATA SCIENCE

The term “data science” was first introduced by William S. Cleveland in 2001 when he published his paper entitled “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics”. It is a plan “to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called ‘data science’ [3].” He proposed to put this new discipline in the context of computer science and data mining. Due to the extraordinary growth of data during the past few decades, data scientists have become one of the most attractive jobs of the 21st century [4]. Undoubtedly there is a growing demand for data science professionals in businesses. The supply of professionals who can work effectively with such big data is limited. A recent study by the McKinsey Global Institute concludes “The United States has led the big data revolution but faces a shortage of the analytical and managerial talent needed to maximize its potential” [17].

Since the supply of data scientists seems to be very limited throughout the world, the following case study aims to assist talent mobility manager to better match talent supply and demand in the area of data science in Thailand. To achieve this goal, we plan to answer the following three research questions:

1. What are the main areas of focus for data science research activities in Thailand?
2. Who are the key experts in the field of data science in Thailand?
3. Are there any forms of association between these key experts?

A. The Data

In Section I, we described the incomplete areas of expertise value in the current Talent Database. When we performed keyword search with query “data science” on the Talent Database, we found zero expert. We expand our search terms to “data mining,” “data analysis,” “statistics,” and other related keywords to data science, we retrieved the number of experts as shown in Table 1.

TABLE 1 - NUMBER OF RETRIEVED EXPERTS FOR DIFFERENT SEARCH QUERY

Search Query	Number of Retrieved Experts
Data Science	0
Data mining	10
Data analysis	6
Statistics	5
Machine learning	2
Database system	3
Text mining	5
Artificial intelligence	7

As we can see from Table 1, the number of retrieved experts from the current Talent Database for each specific area related to data science is quite limited. In order to generate a more complete view of data scientists and their areas of expertise in Thailand, we decided to exploit publications as an alternative source of information. The keywords or subjects indexed in publications may serve as a functional equivalent to an expert's areas of expertise.

We retrieved the publications of 17,863 Thai researchers from the SCOPUS[2] database. Out of 17,863, we were able to retrieve 42,730 publications from 9,649 researchers from the year 2000 onward. Note that some researchers had their names in the TNRR list, but had no publications indexed in SCOPUS. We theorize that these may primarily be local researchers whose research outputs are not formally published, but the reason for these omissions may not be relevant to our inquiry. We proceeded to profile data science R&D activities in Thailand based on these 42,730 publications.

In order to retrieve publications related to the area of data science, we selected publications with keywords such as "data science," "data mining," "data analysis," "database system," "text mining," "information retrieval," "artificial intelligence," and "statistics," among others. We were able to retrieve 1,413 publications related to data science.

B. The Analysis

We proposed to apply bibliometric analysis and text mining to help identify research activities in data science in Thailand. The design for this process (as shown in Fig. 1) is based on the Tech Mining approach proposed by Porter and Cunningham [28]. We first imported the 1,413 publications related to data science into Vantagepoint[3], an analytical text mining software. To profile R&D activities, the program applies "factor analysis" and "principal component analysis" of keywords to help discern sub research areas within "data science." In addition, besides reporting the list of top "n" experts in data science, the program can identify linkages between experts based on their co-authoring relations and based on keywords they share in their publications. The results could then be used to assist talent manager to better match talent demand with the available supply.

C. The Results

1. Data Science Topical Emphases in Thailand

To identify the main foci of research in Thai data science, a map of keywords is created. Fig. 2 presents a high level map of clusters of the 290 most frequently occurring keywords in the "data science" dataset (1,413 records). This clustering is based on a "principal components analysis" to group keywords that often appear together in the records. Different nodes identify different factors or clusters of these highly correlated keywords. Node size reflects the frequency of documents represented by those terms. Placement of nodes is based on VantagePoint's Multidimensional Scaling (MDS) algorithm. Topics that co-occur together will be placed near each other. Connecting lines represent the strength of the association between two clusters, based on a "Path Erasing Algorithm" [38].

The interlink between nodes in the central region of Fig. 2 shows concentration of research on data mining applied to the medical and healthcare industry. Other dominant research areas in data science include facial recognition, image processing, bioinfomatics, text mining and ontology.

2. The Key Data Scientists in Thailand

Once we identify the key research areas as shown in Fig. 2, we can identify who the key researchers in each area are. Table 2 shows a distribution of the number of publications for some of the key research areas from top researchers. This table indicates who the prominent data scientists for each area are. For instance, if industry is looking for a data scientist in image processing techniques applied to facial recognition, talent mobility manager can contact P. Sanguansat from Panyapiwat Institute of Management or S. Marukatat from National Electronics and Computer Technology Center as possible candidates for talent mobility.

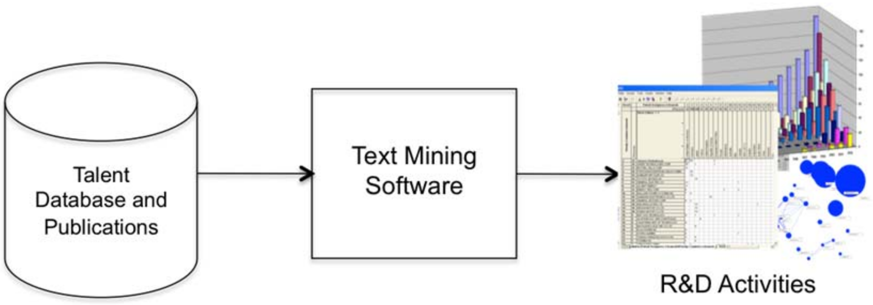

Fig. 1- Bibliometric Analysis and Text Mining Process

[2] SCOPUS is produced by Elsevier; it contains abstracts and citations from journal articles and conference papers from several thousand sources providing coverage of scientific, technical, medical and social sciences fields and arts and humanities.

[3] VantagePoint is a text-mining software for discovering knowledge in search results from patent and literature databases (http://www.thevantagepoint.com/)

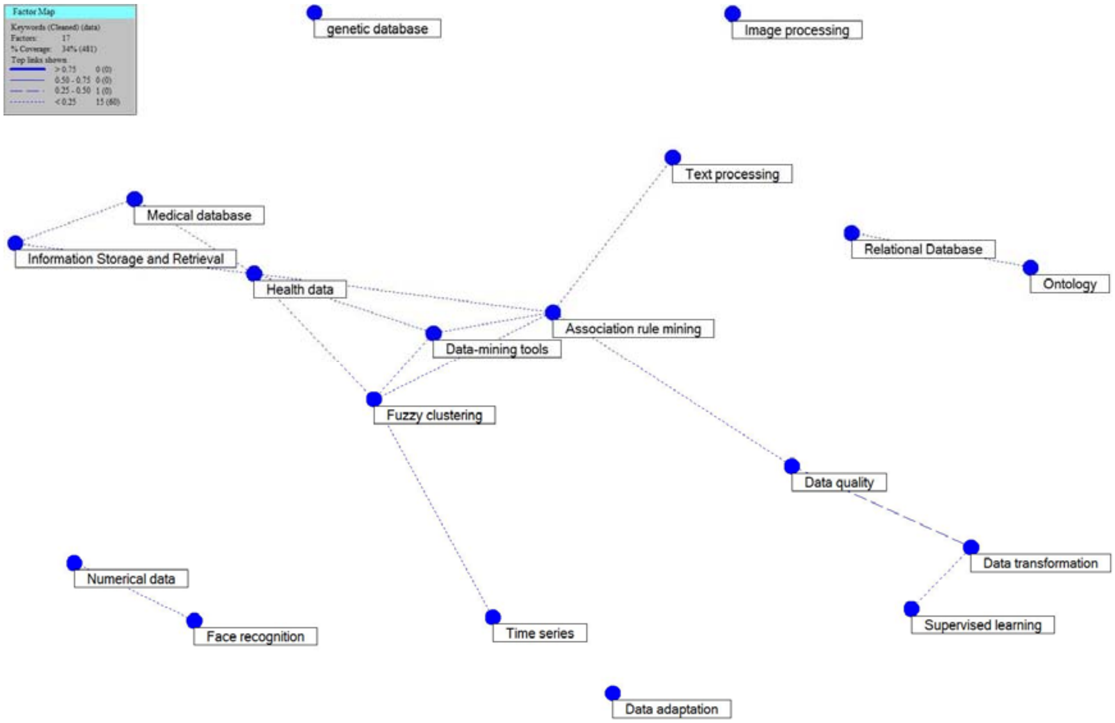

Fig. 2 - Main Foci of Research in Thai Data Science

TABLE 2: DISTRIBUTION OF NUMBER OF PUBLICATIONS FOR SOME OF THE KEY RESEARCH AREAS FROM TOP RESEARCHERS

		# Publications						
#	Researchers	Association rule mining	Ontology	Data-mining tools	Text processing	Facial recognition	Genetic database	Fuzzy clustering
27	Kerdprasop, N	20	1	5		1		4
25	Kerdprasop, K	19	1	4		1		3
19	Ratanamahatana, C.A	7		4	1	2		1
19	Lursinsap, C	2	1	1	1	2	1	3
19	Theeramunkong, T	8	3	1	7	1		1
15	Marukatat, S	2		2	3	5		
12	Kijsirikul, B	2			4	2		
12	Sanguansat, P			2		9		
12	Haruechaiyasak, C	8	3	3	3			4
12	Tongsima, S	1				1	7	
12	Natwichai, J	5	2					
12	Chongstitvatana, P			3	1			1
11	Niennattrakul, V	4		1	1	2		1
11	Jaruskulchai, C			1	1	1		2
11	Assawamakin, A	1				2	6	

3. Relationships Among Key Researchers

Often times we may identify certain candidates for talent mobility who are not available for or interested in joining the project. If we need to locate others, one approach that could help identify other potential researchers is to discover associations among our experts. Typically, associations between researchers are determined by: 1) co-authorship [19, 22] or 2) shared common research topic [6]. Fig. 3 presents “knowledge networks” of top data scientists based on their co-authorship. The more often researchers publish papers together, the stronger their association. Stronger associations are depicted by closer proximity in the diagram.

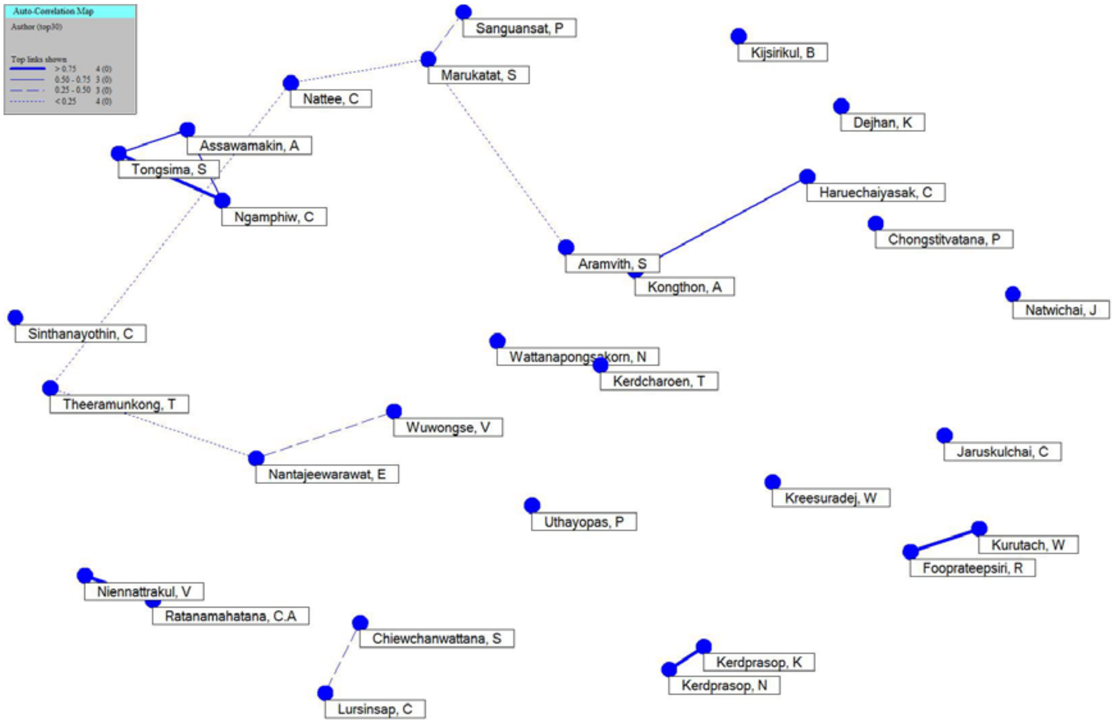

Fig. 3 - Associations of the Top 30 Data Scientists Based on Co-Authorship

Fig. 4 presents another type of association among top data scientists. These associations are based on shared common topics or keywords. The more topics or keywords the experts share in their publications, the greater likelihood that they work in a similar research area, even though they may or may not be co-authoring together.

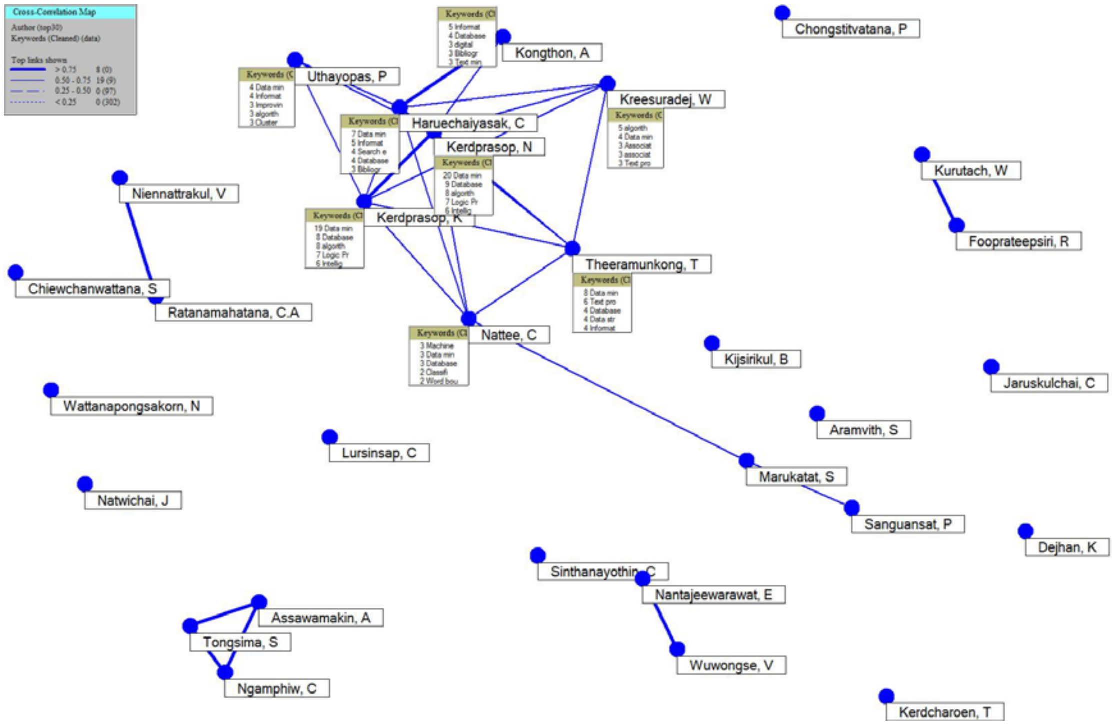

Fig. 4 - Associations of the Top 30 Data Scientists Based on Their Shared Keywords

Suppose from Table 2, a talent mobility manager is able to locate S. Tongsima as an expert in bioinfomatics. Therefore, she is trying to contact him to join a talent mobility program. In the event that S. Tongsima is not available to participate, A. Assawamakin or C. Ngamphiw could be potential candidates because they have written some publications on bioinformatics together (see the upper left corner of Fig. 3).

For another case, suppose a company would like to locate experts in the area of data mining to help them solve some specific problems. A talent mobility manager can use association networks as shown in the upper part of Fig. 4 to identify potential experts in data mining. Top keywords associated with each expert can be identified by examining the "pull-down" boxes (as shown in Fig. 4). For example, C. Haruechaiyasak and T. Theeramunkong are experts in text processing, information retrieval and search engines. Furthermore, N. Kerdprasop and K. Kerdprasop are experts in data reduction, data clustering and association rule mining.

IV. DISCUSSION

In order to mobilize the right talent, possessing the necessary skills, we need to have a comprehensive Talent Database. Setting up a system and inviting all researchers to provide their areas of expertise voluntarily may be overwhelming to the participants and the database administrator. The current Talent Database only supports keyword search for potential candidates. However, a keyword search alone is not sufficient to match the talent demand with the available supply in the database because:

1. Keyword searches do not provide the "big picture" of related sub topics in the areas of interest and
2. Keyword searches provide only a list of researchers with no indication of the associations between them.

For the aforementioned reasons, applying bibliometric analysis and text mining can help a talent mobility manager more accurately determine topical emphases and experts' knowledge networks in their area of research. As shown in the illustrative case of data scientists, the talent mobility manager can first identify the main foci of data science research in Thailand. These areas may include topics such as facial recognition, image processing, bioinfomatics, text mining and ontology. Other major applications of data science in Thailand are in the medical and healthcare sector.

The talent mobility manager can then identify experts and their knowledge networks based on their co-authorship and their shared similar research topics. Co-authoring is a direct association among experts. We may assume that experts who co-author will most likely have similar expertise. Another form of association is based on experts who do not co-author but share the same research topics. Using both derived associations, a talent mobility manager can both expand their target list of experts for some sub-topics within data science and determine which are more likely to possess the target skills and knowledge.

V. CONCLUSIONS

Because the current Talent Database is incomplete and the area of expertise field has been left blank for a great number of researchers, we proposed an approach to extract more areas of expertise from publications. In addition, we applied bibliometric analysis and text mining to provide insights into related sub topics in the areas of interest, as well as to identify experts and their collaborations. Compared to existing keyword matching mechanism, this proposed method could assist a talent mobility manager to better match talent demand from the industry with the available supply of experts. However, even though keywords from publications may be a reasonable proxy for areas of expertise, the Thai government still needs to find an effective approach to encourage experts to register with the Talent Database and keep updating their profile regularly.

REFERENCES

[1] Börner, K., C. Chen and K.W. Boyack, "Visualizing knowledge domains," *Annual review of information science and technology*, 37(1): 179-255, 2003.

[2] Callon, M., J.P. Courtial, W.A. Turner and S. Bauin, "From translations to problematic networks: an introduction to co-word analysis," *Social Science Information*, 22, pp. 191-235, 1983.

[3] Cleveland, W., "Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics," *International Statistical Review*, 69(1), pp. 21–26, 2001.

[4] Davenport, T. and D.J. Patil, "Data Scientist: The Sexiest Job of the 21st Century," *Harvard Business Review*, October 2012 Issue.

[5] Garfield, E., I.H. Sher and R.J. Torpie, The use of citation data in writing the history of science. DTIC Document, 1964.

[6] Gerdsri, N., A. Kongthon and S. Puengrusme, "Discovering the professional communities and social networks of emerging research areas: Use of technology intelligence from bibliometric and text mining analysis" *In Proceedings of the 2012 Portland International Conference on Management of Engineering and Technology (PICMET 2012)*, pp. 114-121, 2012.

[7] Gerdsri, N., A. Kongthon and R. Vatananan, "Mapping the Knowledge Evolution and Professional Network in the Field of Technology Roadmapping (TRM): A Bibliometric Analysis," *Technology Analysis & Strategic Management*, 25(4): 403-422, 2013.

[8] Garner, J., A.L. Poter, N.C. Newman and T.A. Crowl, "Assessing Research Network and Disciplinary Engagement Changes Induced by an NSF Program," *Research Evaluation*, 21(2), pp. 89-104, 2012.

[9] Guthridge, M., A.B. Komm and E. Lawson, "Making talent management a strategic priority," *The McKinsey Quarterly*, January: 49-59, 2008.

[10] Kongthon, A., "A text mining framework for discovering technological intelligence to support science and technology management," Doctoral Dissertation, Georgia Institute of Technology, 2004.

[11] Kostoff, R.N., "Literature-related discovery:introduction and background," *Technological Forecasting & Social Change*, 75(2), pp. 165-185, 2008.

[12] Kostoff, R.N., "Literature-related discovery and innovation – update," *Technological Forecasting & Social Change*, 79(4), pp. 789-800, 2012.

[13] Kostoff, R.N., J.A. Block, J.A. Stump and D. Johnson, "Literature-related discovery: potential treatments for Raynaud's Phenomenon," *Technological Forecasting & Social Change*, 75(2), pp. 203-214, 2008.

[14] Kostoff, R.N., M.B. Briggs and T. Lyons, "Literature-related discovery: potential treatments for multiple sclerosis," *Technological Forecasting & Social Change*, 75(2), pp. 239-255, 2008.

[15] Kostoff, R.N. and U. Patel, “Literature-related discovery and innovation: Chronic kidney disease,” *Technological Forecasting & Social Change*, 91, pp. 341-351, 2015.

[16] Leydesdorff, L. and I. Rafols, “A global map of science based on the ISI subject categories,” *Journal of the American Society for Information Science and Technology*, 60(2), pp. 348-362, 2009.

[17] Manyika, J. and M. Chui, “MBAs can't afford to end their math education with calculus,” *Business Insider*, 2013. http://www.businessinsider.com/why-statistics-is-worth-more-than-calc-2013-3.

[18] McBurney, M. K. and P.L. Novak, “What is bibliometrics and why should you care?” *IEEE Professional Communication Conference (IPCC)*, Portland, Oregon, pp. 108-114, 2002.

[19] Melin, G. and O. Persson, “Studying research collaboration using co-authorships,” *Scientometrics*, 36, pp. 363–377, 1996.

[20] Ministry of Science and Technology, “The Cabinet approves “Talent Mobility Policy” to promote the researchers and the government-scholarship students working in the industrial sectors,” http://www.most.go.th, Access on February 13, 2016.

[21] National Science Technology and Innovation Policy Office, *The Research, Development, and Innovation Survey in Thai Private Sector*, Ministry of Science and Technology, Thailand, 2008. (in Thai language)

[22] Newman, M.E.J., “Coauthorship networks and patters of scientific collaboration,” in *Proc Natl Acad Sci*., 101, pp. 5200-5205, 2004.

[23] OECD, “The global competition for talent: Mobility of the highly skilled,” Paris; OECD Publications, 2008.

[24] Oishi, N., “The Limits of Immigration Policies: The Challenges of Highly Skilled Migration in Japan,” *American Behavioral Scientist*, first published on April 13, 2012 doi:10.1177/0002764212441787.

[25] Oishi, N., “Redefining the “Highly Skilled”: The Points-Based System for Highly Skilled Foreign Professionals in Japan,” *Asian and Pacific Migration Journal*, 23 (4), pp. 421-450, 2014.

[26] Pei, R. and A.L. Porter, “ Profiling leading scientists in nanobiomedical science: interdisciplinarity and potential leading indicators of research directions,” *R&D Management*. 41(3): 288-306, 2011.

[27] Porter, A.L., A. Kongthon and J.C. Lui, “Research profiling: Improving the literature review,” *Scientometrics*. 53(3): 351-370, 2002.

[28] Porter, A.L. and S.W. Cunningham, Tech mining: Exploiting new technologies for competitive advantage. New Jersey: Wiley-Interscience, 2005.

[29] Porter, A.L. and I. Rafols, “Is science becoming more interdisciplinary? Measuring and mapping six research fields over time, *Scientometrics*, 81(3), pp. 719-745, 2009.

[30] Porter, A.L., J. Garner and T.A. Crowl, “The RCN (Research Coordination Network) experiment; Can we build new research networks?, *BioScience*, 62, pp. 282-288, 2012.

[31] Price, D.S., Big science, little science. Columbia University, New York, 1963.

[32] Puteh, F., M. Nor and S.H.N. Zulkifli, “Determinants of employment mobility trend among Malaysian young talents,” *in 2012 IEEE Symposium on Business, Engineering and Industrial Applications (ISBEIA*), pp. 102-107, 2012.

[33] Reiner, C., “Brain competition policy as a new paradigm of regional policy: A European perspective,” *Papers in Regional Science*, 89(2), pp. 449-461, 2010.

[34] Small, H., “Cocitation in the scientific literature: a new measure of the relationship between two documents,” *Journal of the American Society for Information Science*, 24(4), pp. 265-269, 1973.

[35] Wipawin, N. and A. Wanna, “Institutional Repositories in Thai Universities,” *In Proceedings of The Emergence of Digital Libraries – Research and Practices: 16th International Conference on Asia-Pacific Digital Libraries, ICADL 2014*, pp. 385-392, 2014.

[36] World Economic Forum, “Talent Mobility Good Practices – Collaboration at the Core of Driving Economic Growth,” World Economic Forum, Switzerland, 2011.

[37] World, Economic Forum, “The Human Capital Report 2015,” World Economic Forum, Switzerland, 2015.

[38] Zhu, D. and A.L. Porter, “Automated Extraction and Visualization of Information for Technology Intelligence and Forecasting,” *Technological Forecasting and Social Change*, 69, pp. 495-506, 2002.