Indonesian Journal of Electrical Engineering and Computer Science
Vol. 3, No. 1, July 2016, pp. 200 ~ 208
DOI: 10.11591/ijeecs.v3.i1.pages

Technology Map: A Text Mining and Network Analysis Approach

Arif R Hakim[1*], Taufik Djatna[1], Ade Febransyah[2]
[1]Postgraduate Program in Dept. Agro-industrial Technology – Bogor Agricultural University, Kampus IPB Darmaga P.O. Box 220 Bogor, (62-251) 8621974/(62-251) 8621974
[2]Prasetiya Mulya Business School
Jl. RA Kartini (TB Simatupang) Cilandak Barat, Jakarta, (62-21) 7511126/(62-21) 7511128
*Corresponding author, e-mail: hakim.arif.r@gmail.com

Abstract

The Information related to the growth of technology is indispensable for research planning. This information is needed by researcher to determine research topics he/she will contribute. For journal editor, this information is needed to evaluate research paper draft. Unfortunately, the growth of technology is unable to be measured directly. Thus, several methods have been developed to measure this growth using patent and journal as its data. Most of the methods were for document which is written in English language. In this work we focused on the development of technology map from Indonesian journal article and its measurement which is unavailable yet. Our method can be used to provide information of the growth of Indonesian technology based on journal article by using text mining and network analysis. For a study case, we used collection of journal article issued in "Jurnal Teknologi Agroindustri". The method can be used to identify correlation between researches in the collection. Furthermore, path of the research growth can be developed for research group in the map. The method can also be used to measure popularity, importance, affinity of the researches to a research group, research type (breakthrough or incremental) and to measure retention time and saturation of a research group.

Keywords: *technology map, text mining, network analysis, technology measurement, Indonesian journal*

1. Introduction

The Information related to the growth of technology development is indispensable for research planning. Among of the stakeholders who need such this information are researcher and journal editor. For researcher, this information can be used to assess topic trend in particular research area, thus he knows in which topics he will contribute. On the other side, journal editor needs this information to conduct initial evaluation of received draft. Unfortunately, activities in technology development is unable to be measured directly. Thus, researchers in technology planning defined indicators to measure the activities [1]. There are two main indicators which are used to measure the activities that are journal and patent. These indicators have positive correlation to the development of science and technology [2].

Three types of data can be used in journal and patent analysis, there are metadata, bibliography, and document content. We can select one of these type of data or combining it to conduct deeper analysis [1]. We can also combine it with other data such as product catalogue or product specification for specific needs [3]. Analysis of these technological documents resulting in important specific information of technology development, depent on the approach. Some of the information that can be obtained are future alternatives technology [4], emerging technology [5-8], new technology [9], technology roadmap and trend [3, 5], [9-11], important technology [9, 12], path of technology development [12], technology mapping [13], technology evaluation [13], cluster of technology [14], innovation pattern in global or regional scope [15], promising technology [16-18], and patent creation guidance [19]. The analysis can be fully numerical or can be enhanced with data visualization. By adding data visualization, analyst has more point of view to the data and he is able to observe global pattern of the data.

Analysis of technological document is devided into network-based analysis and keyword-based analysis. Both methods have their strength and weakness. In network-based, internal information cannot be analyzed, while in keyword-based analysis the correlation

Received February 2, 2016; Revised May 25, 2016; Accepted June 10, 2016

between document is missing [10]. Thus the combination of these methods are irresistible. An approach to combine these methods is by extracting keywords in the document and creating the correlation between documents using the keywords [3, 10], [20-21]. The output of this approach is documents relationship, which is able to be visualized in graph. As the relationship is visualized as a graph, it is possible to conduct network analysis to point out the important information from the technology graph to model the activities in technology growth. This approach has advantages in the ease of visualization and in its ability to identify internal relationship between documents [21].

As we model the research development as graph, it enables network analysis to find important information from the network. Network analysis can be used to find information in abundant of problem as long as the problem can be modeled into graph, such as in social media. This approach can be used to identify spammer and to determine the suitable groups for spreading advertisement messages in twitter [22]. It also can be used to find important information from Facebook such as important person in a community and tracking the change of the people role in the community [23]. Network analysis on a graph is conducted by measuring properties of nodes and vertice in the graph, for example by simply counting all nodes connected to a node, we know the popularity of a node compared to other nodes in the graph. Several measurements are available in network analysis such as degree, betweenness centrality, closeness centrality, etc. In this work we applied these measurements to the network of technology.

All the methods of technological analysis, which are described above, were used for documents which are written in English. In this work, we focused on the development of technology map from journal articles which are written in “Bahasa Indonesia” using text mining approach and evaluating researches in the map using network analysis to combine network-based analysis and keyword-based analysis. We also combine content and metadata of the journal articles for deepest information result. In the last but not least, we provide visualization of the developed network, we named it technology map, to enables tracking of the technology growth and evaluation of the researches for technological research planning.

2. Research Method

2.1. Technology Map Foundations

In order to develop technology map, we represented path of technology growth as a graph G which consists of set of researches, represented by set of vertices $V = \{v_1, v_2, v_3, ..., v_i\}$ and correlation of the researches, represented by set of edges $E = \{v_1v_2, v_1v_3, ..., v_iv_j\}$, thus $G = \{V, E\}$ (see Figure 1). In this work, we used journal articles (writtten in “Bahasa Indoensia”) which are represented by collection of journal articles $D = \{d_1, d_2, d_3, ..., d_k\}$ while d_k is individual research article. To model the collection into graph, we assumed two important assumptions. First, research topic T of a journal article d_k is represented by set of words W and its frequency F, thus $T_{dk} = \{W, F\}$ for each document. Second, if we aggregate T_{dk} in D, then we get n-dimensional space R^n, n represents number of unique W which is exist in D. Position of d_k in the R^n is determined by T_{dk}. Thus, graph G of D can be developed by connecting d_k in R_n based on its T_{dk}.

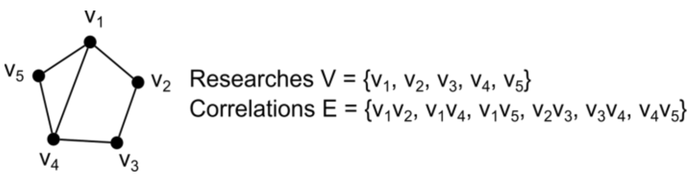

Figure 1. A Graph Represents Research Network (adapted from [24])

2.2. Data

In this work, we used collection of journal articles which are written in “Bahasa Indonesia”. There were 196 journal articles in agro-industrial technology issued in “Jurnal Teknologi Industri Pertanian” from 2004 to 2013. Since we focused on journal which is written in

"Bahasa Indonesia", we excluded twelve journals which are written in English. These data sets were retrieved from http://journal.ipb.ac.id/index.php/jurnaltin/issue/archive. Almost all of the article consist of title, author(s) of the article, the author(s)' institution, abstract, keywords, article content, and bibliography. Such types of information were equally processed.

2.3. Generating Technology Map using Text Mining

We adopted the approach conducted by Yoon dan Park [21] which is most flexible approach to develop technology map with several modifications to compromise Indonesian journal articles and enable tracking of technology growth. We also made a revision of the equation used by Yoon and Park [21] to measure correlation between researches. We used stop words dictionary for stop words removal, root words dictionary and stemming method [25] for "Bahasa Indonesia". Stop words removal was to improve the result of keywords extraction process [26], while stemming was to enhance the accuracy of correlation detection [27].

Collected journal articles were in portable document format (pdf), thus we converted the documents into plain text using pdf2text before were being tokenized. Tokenization was done using Regex by identifiying all word without including white space and punctuation mark. All stop word and affix in the Tokenized text was cleared. Identification of topic of science articles was done by evaluating words in the documents based on TF-IDF scoring (Equation 1), where $TF\text{-}IDF^i_k$ is TF-IDF score of word k in document i, f^i_k is frequency of word k in document i, N is the amount of document in the collection, and d_k is the amount of document which consists of word k. We accepted 80% words with highest scored to avoid over removal if the document is shorter.

$$TF\text{-}IDF^i_k = f^i_k \times \log\frac{N}{d_k} \tag{1}$$

$$association^{ij} = 1 - \sqrt{\frac{\sum_{k=1}^{K}(f^i_k - f^j_k)}{K}} \tag{2}$$

Association between researches in the collection was measured using Equation 2. The equation was revised from [21] in order to get value between 0-1. Notation f^i_k in the equation is frequency of word k in document i, K is the amount of evaluated word. Based on our observation, score of 0.07 was the lowest value of identifiable correlation. Clauset-Newman-Moore method [28] was used to cluster the research based on its correlation to build group of research. The clustered researches then was mapped using NodeXL [29]. Framework of the method is shown in Figure 2.

2.3. Evaluation of Technology in Technology Map using Network Analysis

As the map was built, the technologies in map can be evaluated using network analysis. Several measurements which are used are shown in Tabel 1. In addition, we propose three measurements for evaluating researches in the map, i.e. type of research, research group retention time, and reseach group saturation index.

Type of research were used to identify if a research belongs to breakthrough or incremental research. Breakthrough research is a research without prior research in the same research group. Otherwise, incremental research is a research which follows prior research in the same research group. The type of a research i was determined using Equation 3, where I_{ij} is a binary value which has value 1 if research i is older than research j for all j in J and 0 if research i is newer than research j for all j in J, while i and j are connected research. A breakthrough research has type value of 1, while incremental research has type value of 0.

$$type_i = \prod_{j=1}^{J} I_{ij} \tag{3}$$

$$retention_g = 1 + max_{i \in g}(y_i) - min_{i \in g}(y_i) \tag{4}$$

$$saturation_g = \frac{\mu_g}{retention_g} \tag{5}$$

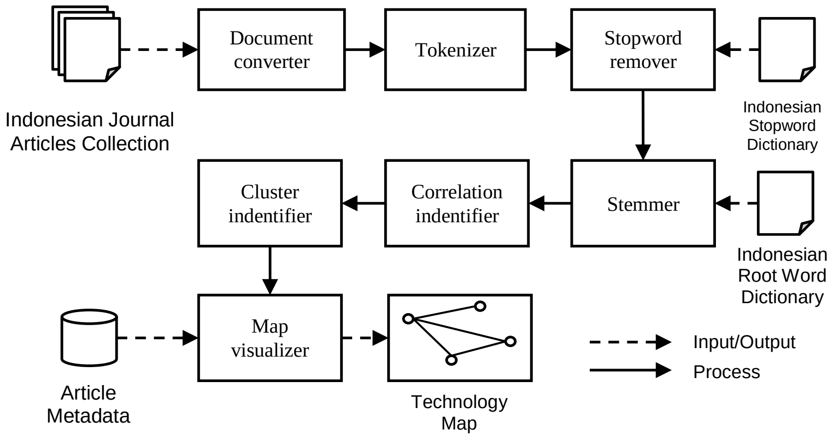

Figure 2. Framework for Developing Technology Map

Retention time indicates the time needed by a research group to reach its vacuum, in the other words there is no researchers are interested to the research area. Retention time of a research group *g* was determined using Equation 4, where $max(y_i)$ is the year of newest research is conducted within the group and $min(y_i)$ is the oldest research is conducted within the group.

Saturation index indicates the saturation level of a research group. A group of research is more saturated if it has more research in shorter retention time, *vice versa*. Saturation index was calculated using Equation 5, where μ_g is the number of member in a research group *g*.

Table 1. Measurements used in Evaluation of Technology in Technology Map

Measurement	Description
Degree [28]	Degree was used to measuring popularity of the research. Popular research has more correlation to others researches. Higher degree means that a research more popular than lower degree.
Betweennes Centrality [30]	Betweeness centrality was used to measuring importance of a research in the path of technology growth. Higher betweeness centrality means that a research more important in the path of technology growth than lower betweeness centrality.
Closeness Centrality [31]	Closeness centrality was used to measuring the strength of affinity of a research to its group. Higher closeness centrality means that a research has more strong correlation than lower closeness centrality.

Table 2. Association Values Between Several Documents

Title of Journal article 1	Title of Journal article 2	Association value
pengaruh nisbah pereaksi (lignin eupcalyptus – natrium bisulfit) dan ph awal reaksi sulfonasi terhadap karakteristik natrium lignosulfonat	pembuatan natrium lignosulfonat berbahan dasar lignin isolat tandan kosong kelapa sawit: identifikasi, dan uji kinerjanya sebagai bahan pendispersi	0.5528
kajian penggunaan lidah buaya (aloe vera) dan bee pollen pada pembuatan sabun opaque	kajian pengaruh penambahan lidah buaya (aloe vera) terhadap mutu sabun transparan	0.2929
model simulasi dan rancang bangun kapasitas usaha penyulingan minyak nilam	limbah penyulingan sereh wangi dan nilam sebagai insektisida pengusir lalat rumah (musca domestica)	0.0742
karakteristik mutu sop daun torbangun (coleus amboinicus lour) dalam kemasan kaleng dan perhitungan total migrasi bahan kemasan	penentuan komposisi atmosfir untuk penyimpanan bawang daun rajangan	0.0691
deasidifikasi dan dekolorasi minyak jarak pagar (jatropha curcas l.) dengan menggunakan membran mikrofiltrasi	permodelan matematika ekstraksi oleoresin temulawak (curcuma xanthorrizha roxb) dengan karbondioksida superkritis dan co-solvent etanol menggunakan shrinking core model	0.0000

3. Results and Analysis

By using the framework previously described, a technology map can be developed as shown in Figure 3. The map was drawn based on correlation of researches in the collection. In the map, researches which have association to other research are drawn as graph in the map, while the researches without association are shown in the top of the maps as a black dot. From the data we found that 64 of 184 researches have one or more association, while 120 researches have no association (Table 2 shows comparison of several researches with its association value). There are 19 research groups consist of 2 to 11 members which are shown in different color. We visualize the correlation of two researches by a line. Thicker line means higher association value. This makes identification of relevant research in the collection can be conducted rapidly.

Researcher is able to monitor annual growth of each research group. Figure 4 is an example of this feature representing researches in bio oil usage during 2009 to 2012 (research titles of group member are shown in Table 3). This group consists of several research topics, that are rubber seed oil usage for leather tanning, and biodiesel from jatropha oil, rubber seed oil, and palm oil. In 2009 until 2010, the group was only consisting of three researches focused on rubber seed oil for leather tanning. The group was extended in 2011 as three of researches in bio-fuel production from palm oil were conducted. These three of researches have no correlation to the prior researches because the prior researches uses rubber seed oil for leather tanning, not for biodiesel. After 2012 these two group were connected because a research which is focused on the usage of rubber seed oil and palm oil as raw material in biodiesel production was conducted (reseach 183). Research 183 was an important bridging research between two previous research topic. This finding was indicated by the result of analysis of the research group as shown in Table 4.

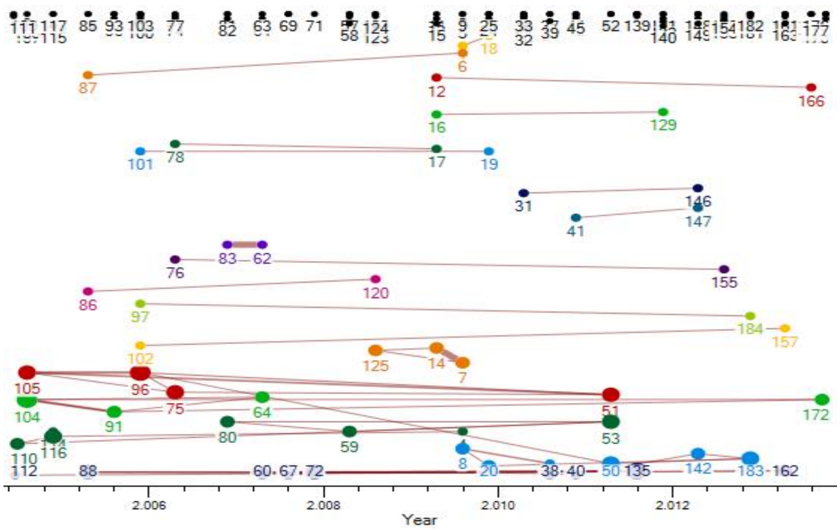

Figure 3. Technology Map in Agro-Industrial Technology Research Area

Table 3. Research Group in Rubber Seed Oil an Palm Oil for Leather Tanning and Biodiesel

ID	Journal article Title
8	optimisasi pengeringan biji karet (hevea brasiliensis) pada ekstraksi minyak biji karet untuk penyamakan kulit
20	potensi pemanfaatan biji karet (hevea brasiliansis muell.arg) sebagai sumber energi alternatif biokerosin
36	penentuan kondisi terbaik pengempaan dalam produksi minyak biji karet (hevea brasiliensis) untuk penyamakan kulit
47	optimasi proses sintesis gliserol tert-butil eter (gtbe) sebagai aditif biodiesel
49	transesterifikasi in situ biji jarak pagar: pengaruh jenis pereaksi, kecepatan pengadukan dan suhu reaksi terhadap rendemen dan kualitas biodiesel
50	karakteristik biodiesel dan blending biodiesel dari oil losses limbah cair pabrik minyak kelapa sawit
142	optimalisasi kinerja pembuatan dan peningkatan kualitas biodisel dari fraksi minyak limbah cair pengolahan kelapa sawit dengan memanfaatkan gelombang ultrasonik
183	pembuatan biodiesel biji karet dan biodiesel sawit dengan instrumen ultrasonik serta karakteristik campurannya

In Table 4, we show the result of analysis in research group of bio oil usage from the previous example. There are four of measurements was conducted to the researches in the group. We can identify the most popular research from the group are research 49 and research 183 (highest degree in the example). By considering its betweenness centrality, research 183 is also the most importance research for the growth path of the group. When this research is removed from the path, the group divided into three groups that are research about rubber seed oil usage, palm oil usage, and additive usage in bio-fuel reaction. This indicator shows that the research has important role in bridging between researches in the research group as stated before. On the other hand, closeness centrality shows that research 49 and research 183 are the central of the research group. It means that the topic of these researches are the most influencing topic compared to other researches. Using these measurements, we can identify the most important research in a group, that are research 49 and 183.

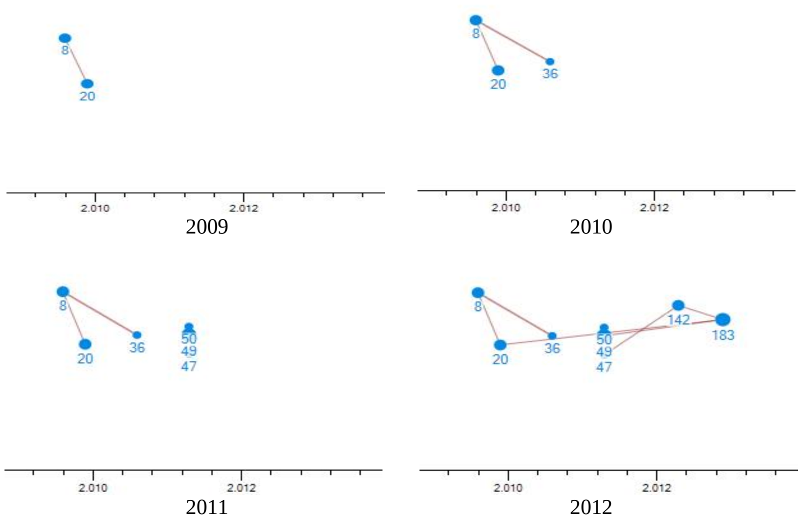

Figure 4. Development of Research in Bio Oil usage during 2009 to 2012

Other research analysis was determination of research type. In the group of bio oil usage, there are four breakthrough researches and four incremental researches. Research 8 was a breakthrough research because it was the first research in this area, while research 47, 49 and 50 became breakthrough since it was the first research focusing on bio-diesel production and the researches was conducted at the same time.

Table 4. Measurement of Researches in Bio Oil Usage

Measurement	Document							
	8	20	36	47	49	50	142	183
Degree	2	2	1	1	3	1	2	3
Betweenness centrality	10	18	0	0	34	0	10	36
Closeness centrality	0.027	0.034	0.021	0.024	0.034	0.030	0.032	0.043
Research type	br	in	in	br	br	br	in	in

The result of measurement of retention time and saturation index of the research article collection is shown in Table 5. There are 19 groups which have different retention time, different amount of member and different saturation index. From the collection of journal article, highest retention time is 10 years and the lowest is 1 year. The highest saturation index in the collection is 2.000 and the lowest is 0.222. On of the group which have the highest saturation index was group 19. It means that there are two new researches in the group every year. Group number 2 is also have the same saturation index, althought group 2 is more active than group 19 based on its retention time and amount of its member. Compared to the other groups, this groups are the most active. Thus, the researcher in this research area have to be more active to synchronize with the growth of technology in the group.

4. Evaluation

We evaluated the result of the approach by comparing it to the data which are manualy labeled. We conducted 350 manual identification of correlation between journal articles in the collection (two articles were compared for each comparison). The result of the evaluation is shown in Figure 5. Based on the evaluation, our approach was able to identify 63% of the total correlation in the sample with 98% of the identification was correct.

Table 5. Result of Research Area Analysis in Agro-Industrial Technology

Group	Retention Time	Member	Saturation index
1	10	11	1.1000
2	4	8	2.0000
3	8	7	0.8750
4	10	5	0.5000
5	8	4	0.5000
6	2	3	1.4999
7	9	2	0.2222
8	8	2	0.2500
9	4	2	0.5000
10	7	2	0.2857
11	2	2	1.0000
12	3	2	0.6667
13	3	2	0.6667
14	5	2	0.4000
15	4	2	0.5000
16	3	2	0.6667
17	5	2	0.4000
18	5	2	0.4000
19	1	2	2.0000

		Prediction	
		Correlated	Not Correlated
Actual	Correlated	50	29
	Not Correlated	1	270

Precision = 0.9804
Recall = 0.6329
F-Measure = 0.7692

Figure 5. Confusion Matrix of Identification of Research Correlation

5. Conclusion

Technology map in this work was developed using text mining approach which consists of seven steps that are document conversion, document tokenization, stopword removal, stemming, association identification, clustering, and graph visualization. By adding metadata such as issued year as research property, it enables monitoring function through the maps. Several measurements, such as degree, betweenness centrality, closeness centrality, type of research, retention time and saturation index enable evaluation of both individual research and research group in the map. The maps can also be used to monitor path of research development and predict potential research area to be explored by selecting unsaturated research area. The defined measurements in this work are useful to evaluate researches within the maps by comparing between the researches in the maps. We found that research area in agro-industrial technology is wide enough to be explored. Based on the evaluation, our approach was able to identify 63% of the total correlation in the collection, with 98% of the detection was valid. This work was an initial step which ia can be used as basis for more complex approach for Indonesian journal article.

References

[1] Barberg I. Patents and the Measurement of Technological Change: A Survey of the Literature. *Research Policy*. 1987; 16: 131-141.
[2] Narin F. Patent Bibliometrics. *Scientometrics*. 1994; 30(1): 147-155.
[3] Yoon B, Phaal R, Probert D. Morphology Analysis for Technology Roadmapping: Application of Text Mining. *R&D Management*. 2008; 38(1): 51-68.
[4] Anderson TR, Daim TU, Kim J. Technology Forecasting for Wireless Communication. *Technovation*. 2008; 28(9): 602-614.
[5] Huang L, Zhang Y, Guo Y, Zhu D, Porter AL. Four Dimensional Science and Technology Planning: A New Approach Based on Bibliometrics and Technology Roadmapping. *Technological Forecasting and Social Change*. 2014; 81: 39-48.
[6] Kim J, Hwang M, Jeong DH, Jung H. Technology Trends Analysis and Forecasting Application Based on Decision Tree and Statistical Feature Analysis. *Expert Systems with Applications*. 2012; 39(16): 12618-12625.
[7] Kajikawa Y, Yoshikawa J, Takeda Y, Matsushima K. Tracking Emerging Technologies in Energy Research: Toward a Roadmap for Sustainable Energy. *Technological Forecasting & Social Change*. 2008; 75: 771-782.
[8] Kim YG, Suh JH, Park SC. Visualization of Patent Analysis for Emerging Technology. *Expert Systems with Applications*. 2008; 34: 1804-1812.
[9] Lee S, Yoon B, Lee C, Park J. Business Planning Based on Technological Capabilities: Patent Analysis for Technology-driven Roadmapping. *Technological Forecasting and Social Change*. 2009; 76(6): 769-786.
[10] Choi J, Hwang YS. Patent Keyword Network Analysis for improving Technology Development Efficiency. *Technological Forecasting and Social Change*. 2014; 83: 170-182.
[11] Jeong Y, Yoon B. Development of Patent Roadmap Based on Technology Roadmap by Analyzing Patterns of Patent Development. *Technovation*. 2015; 39: 37-52.
[12] Choi C, Park Y. Monitoring the Organic Structure of Technology Based on the Patent Development Paths. *Technological Forecasting and Social Change*. 2009; 76(6): 754-768.
[13] Lee PC, Su HN, Wu FS. Quantitative Mapping of Patented Technology the Case of Electrical Conducting Polymer Nanocomposite. *Technological Forecasting and Social Change*. 2010; 77(3): 466-478.
[14] Dereli T, Baykasoğlu A, Durmuşoğlu A, Durmuşoğlu ZD. Enhancing Technology Clustering Through Heuristics by Using Patent Counts. *Expert Systems with Applications*. 2011; 38(12): 15383-15391.
[15] Lee S, Lee HJ, Yoon B. Modeling and Analyzing Technology Innovation in the Energy Sector: Patent-Based HMM Approach. *Computers & Industrial Engineering*. 2013; 63(3): 564-577.
[16] Dereli T, Altun K. A Novel Approach for Assessment of Candidate Technologies with Respect to Their Innovation Potentials: Quick Innovation Intelligence Process. *Expert Systems with Applications*. 2013; 40(3): 881-891.
[17] Park H, Ree JJ, Kim K. Identification of Promising Patents for Technology Transfers using TRIZ Evolution Trends. *Expert Systems with Applications*. 2013; 40(2): 736-743.
[18] Yoon B, Park I, Coh BY. Exploring Technological Opportunities by Linking Technology and Products: Application of Morphology Analysis and Text Mining. *Technological Forecasting and Social Change*. 2014; 86: 287-303.
[19] Jeong C, Kim K. Creating Patents on the New Technology using Analogy-Based Patent Mining. *Expert Systems with Applications*. 2014; 41(8): 3605-3614.
[20] Tseng YH, Lin CJ, Lin YI. Text Mining Techniques for Patent Analysis. *Information Processing and Management*. 2007; 43: 1216-1247.
[21] Yoon B, Park Y. A Text-Mining-Based Patent Network: Analytical Tool for High-Technology Trend. *Journal of High Technology Management Research*. 2004; 15: 37-50.
[22] Barash V, Golder S. Twitter: Conversation, Entertainment, and Information, All in One Network!. In: Hansen D, Shneiderman B, Smith MA. *Editors*. Analyzing Social Media Networks with NODEXL: Insights from a Connected World. Burlington: Elsevier Inc. 2011: 143-164.
[23] Hogan B. Visualizing and Interpreting Facebook Networks. In: Hansen D, Shneiderman B, Smith MA. *Editors*. Analyzing Social Media Networks with NODEXL: Insights from a Connected World. Burlington: Elsevier Inc. 2011: 165-179.

[24] West DB. Introduction to Graph Theory. Second Edition. Delhi-India: Pearson Education. 2002.
[25] Asian J, Williams HE, Tahaghoghi SMM. *Stemming Indonesian.* Proceedings of the Twenty-Eighth Australasian Conference on Computer Science, Australian Computer Society, Inc. 2008; 38: 307-314.
[26] Na D, Xu C. Automatically Generation and Evaluation of Stop Words List for Chinese Patents. *Telkomnika*. 2015; 13(4): 1414-1421.
[27] Mardiana T, Adji TB, Hidayah I. Stemming Influence on Similarity Detection of Abstract Written in Indonesia. *Telkomnika*. 2016; 14(1): 219-227.
[28] Clauset A, Newman MEJ, Moore C, Finding Community Structure in Very Large Networks. *Physical Review*. 2004; 70(6): 066-111.
[29] Hansen D, Shneiderman B, Smith MA. Analyzing social media networks with NODEXL: Insights from a Connected World. Burlington: Elsevier Inc. 2011: 69-78.
[30] Newman MEJ. A Measure of Betweenness Centrality Based on Random Walks. *Social Networks*. 2005; 27(1): 39-54.
[31] Okamoto K, Chen W, Li XY. *Ranking of Closeness Centrality for Large-Scale Social Networks.* Frontiers in Algorithmics Second International Workshop, FAW 2008 June 19-21. Changsha, China. 2008: 186-195.