International Journal of Advanced Culture Technology Vol.10 No.4 361-368 (2022)
DOI https://doi.org/10.17703/IJACT.2022.10.4.361

IJACT 22-12-42

Rearch of Late Adolcent Activity based on Using Big Data Analysis

[1]Hye-Sun Lee

[1]Prof., Dept. of Occupational Therapy, Gwangju Women's Univ
lauren52@naver.com

Abstract

This study seeks to determine the research trend of late adolescents by utilizing big data. Also, seek for research trends related to activity participation, treatment, and mediation to provide academic implications. For this process, gathered 1.000 academic papers and used TF-IDF analysis method, and the topic modeling based on co-occurrence word network analysis method LDA (Latent Dirichlet Allocation) to analyze. In conclusion this study conducted analysis of activity participation, treatment, and mediation of late adolescents by TF-IDF analysis method, co-occurrence word network analysis method, and topic modeling analysis based on LDA(Latent Dirichlet Allocation). The results were proposed through visualization, and carries significance as this study analyzed activity, treatment, mediation factors of late adolescents, and provides new analysis methods to figure out the basic materials of activity participation trends, treatment, and mediation of late adolescents.

Keywords: *Late Adolcent, Participation, Activity, Big Data*

1. INTRODUCTION

Mainly found in industrialized countries, the new emerging adulthoods show factors of identity seeking, insecurity, self-centered, feeling in the midst of youth and adult, sensing extended possibilities for the upcoming future. Almost all of these youth receive high education, and have a life cycle of getting married and enter parenthood at around 30 years of age [1]. What are the things late adolescents can do to successfully transition into an adult? Opportunities in activities including education are critical for late adolescents to successfully transition into later stages of life [2]. The World Health Organization's International Classification of Functioning, Disability and Health (ICF) defines health as comprehensive meaning including body function and structure, activity, participation. Participation has been associated with important aspects of daily living such as mobility, social relationships, and activities related to work or school [3]. ICF emphasises how disabilities affect an individual's daily life and participation beyond the meaning of his/her physical and functional deficits [4]. That is, participation in life situations is an important element in determining health and disabilities. The concept of participation is central in habilitation and rehabilitation [5]. Consistent participation or engagement in purposeful and meaningful activities, such as employment, education, and leisure, have been shown to have a positive influence on health, well-being, and subjective life satisfaction [6]. Occupational therapy is a health profession that has a core value of enabling client-centered choice of activities. Occupational therapists prioritize evaluations centered on overall occupations rather than assessments focused on the

Manuscript received: November 26, 2022 / revised: December 4, 2022 / accepted: December 9, 2022
Corresponding Author: insight7@ kwu.ac.kr
Professor, Dept. of Occupational Therapy, Gwang-ju Women's Univ., Korea

physical, emotional, and functional problems experienced by children with disability and they assume client-centered participation as their treatment goal [7]. Due to categorizations of subject & research method and quantity, there were limitations to systematic analysis to understand main keywords and its connections, despite the active research of activity participations of late adolescents until today [8]. This study structures core topics by utilizing text mining and social network analysis to compensate these limitations, and also visualizes network structures of main keywords and scientifically analyze each research directions and its relations to identify research trends of late adolescents.

2. EXPERIMENTS

2.1 Big Data Analysis System

2.1.1 Data Analysis Process

This study proceeds to Gather data to analyze recognition for 'late adolescents' → Data preprocessing → Data Analysis process. Especially for data analysis, text mining analysis methods such as term frequency, TF-IDF, network analysis, LDA analysis to derive key insight for 'late adolescents'.

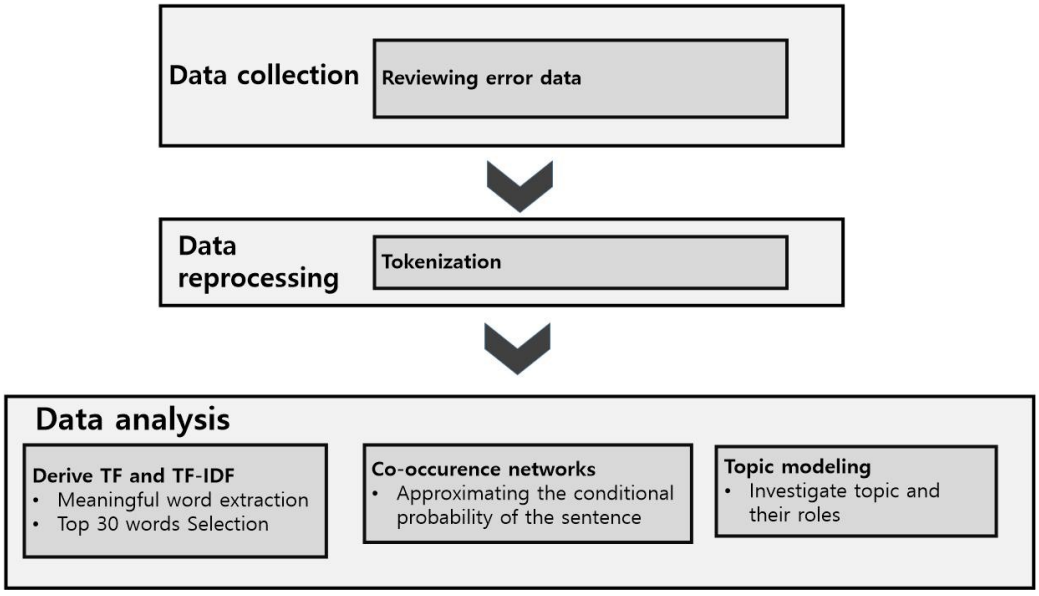

Figure 1. Big Data analysis process

2.1.2 Data Gathering

To gather data, web crawling of academic papers on Google for 'late adolescent' keywords were used. The period of gathering was from January 1st of 2022 to October 1st of 2022. Out of the data gathered, context that were unrelated and of little relevance were excluded from the study. As a result, a total of 1,000 academic papers were gathered, and the study was based on its Abstracts.

2.1.3 Data Preprocessing and Analysis

Study based on text mining method first requires data preprocessing of gathered data for refining and to process data in order for it be easily analyzed. This indicates that language used by people need to be processed

so that computers can understand it [10]. Also, irrelevant data is to be separated and processed to be in accordance with data analysis method and purpose. Data preprocessing was implemented to subdivide morphemes and eliminate stop words to analyze gathered dynamic SQL text data [11]. Review text resources went through morphemes analysis by utilizing NLP sectors' widely used Python Konlpy package's Mecab, and then followed up by tokenizer implementation to extract main nouns, adjectives, and verbs. Afterwards, irrelevant words and phrases, low-relevant numbers, special symbols, and punctuation marks, unrecognizable words were deleted, and researcher repeatedly went over more than 3 times to modify words with postpositions.

2.2 Big Data Analysis Method

2.2.1 Term Frequency and TF-IDF Analysis

Term Frequency is a method to define importance based on the counts of specific terms once the extracted words from preprocessing has been prioritized. In regards to text mining analysis, TF is the most universal analysis method widely used on a regular basis. It is the most basic analysis method to help understand the flow of data during text mining method. This is to show the number of specific words in a document, and the bigger the number of this value, the more often this word is used in a document [12]. Words that are in high frequently usage are normally inherently implying the study topics and can act as a key word [13]. TF-IDF is mostly utilized as morpheme analysis as a statistical numerical data indicating how much of importance a specific word has in a document where groups of documents are composed as one [14]. TF-IDF value increases when some word is not appeared in many documents but are highly mentioned in specific documents. Therefore, if the TF-IDF value is high, there is likely possibility that it carries key significance This study yields TF from word extraction after text preprocessing. Also, this study derives TF and TF-IDF of upper 30 words.

2.2.2 Network Analysis and Centrality Analysis

Network analysis is a method to establish and analyze associative word network by extracting words that appeared simultaneously with specific terms [15]. This method is mainly used to save time and effort for extraction when the size of the data is large. It marks words as node and word connectivity as Edge, and interpret analysis targets by co-occurrence based on these Edges, and understand the structural characteristics of these created networks [16]. It can also be considered connected in words that have high co-occurrence ratio in analyzed networks, and also expressed closely, in Edge terms strongly compared to highly occurred words individually. This study uses Python's network module to calculate weight between highly co-occurred words, and visualizes graphs by calculating weight between the words. Centrality analysis is an index of how specific words are important in word networks, and most representative analysis method is degree centrality [17]. Degree centrality indicates the ratio of nodes that are in connection to real count of nodes and this study uses Python's network module to analyze the Degree centrality.

2.2.3 Topic Modeling

This study derived appropriate topic title which enabled classification of data related to 'late adolescents' using topic modeling method based on LDA(Latent Dirichlet Allocation). LDA is a method to cluster relevant subjects and mainly utilized to analyze unstructured text [18]. This study uses Python's gensim and sklearn

library and proceeded the topic modeling. Selection of appropriate number of topic need to avoid range duplication, and utilized Coherence value and perplexity value to choose the most optimal number of topic.

3. RESULTS

3.1 TF and TF-IDF

After analysis of TF related to 'Late adolescents', 'child(729)' was the highest of ranks, and was followed by 'adult(672)', 'intervention(651)', 'health(640)', 'activity(618)' (Figure 2-a). Words collected apart from top 10 words such as 'health', 'activity', 'participation', etc were deduced related to participation of adolescents, and top 20 words were deducted related to adolescents' therapy such as 'measure', 'therapy, 'symptom'. Also, research words such as parental support such as 'parent', 'support', etc. (Table 1) .TF-IDF analysis of 'late adolescents' shows, 'child(25.59666)' ranked highest, and followed by 'adult(22.45008)', 'intervention(22.12035)', 'health(21.06596)', 'activity(20.02625)'. As similar to TF, similar words were ranked high (Figure 2-b). Besides these words, activity related words such as 'activity(22.45008)', 'participation(17.58533)', 'disorder(16.55083)', ranked top 10, and treatment related words such as 'cancer(15.83276)', 'therapy(15.16334)', 'symptom(11.52597)', 'care(10.93194)' were ranked within 10 to 30 (Table 1).

Table1. TF and TF-IDF of late adolescent

	TF		TF-IDF	
Ranking	**Words**	**Frequency**	**Words**	**Frequency**
1	child	729	child	25.59666
2	adult	672	adult	22.45008
3	intervention	651	intervention	22.12035
4	health	640	health	21.06596
5	activity	618	activity	20.02625
6	age	483	participation	17.58533
7	adolescent	453	adolescent	17.43894
8	life	444	disorder	16.55083
9	year	436	cancer	15.83276
10	participation	430	therapy	15.16334
11	disorder	418	life	15.00022
12	transition	364	age	13.89778
13	self	347	transition	13.17864
14	measure	346	year	13.12177
15	therapy	335	youth	12.56131
16	group	331	qol	12.25751
17	outcome	318	self	12.02102
18	symptom	281	outcome	11.60595
19	qol	280	group	11.58894
20	care	270	symptom	11.52597
21	cancer	269	performance	11.48154
22	assessment	254	measure	11.19203
23	youth	251	care	10.93194
24	time	248	assessment	10.445
25	use	245	parent	10.13155

26	support	241	autism	9.975104
27	performance	236	asd	9.947586
28	level	232	skill	9.841532
29	parent	228	use	9.82828
30	treatment	219	disability	9.775628

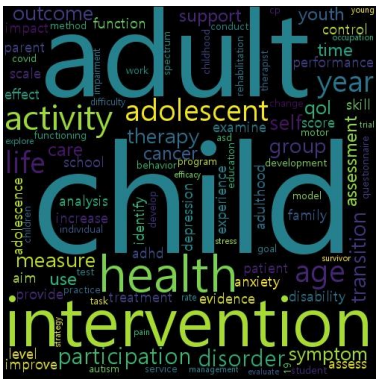

Figure 2- a. late adolescent TF

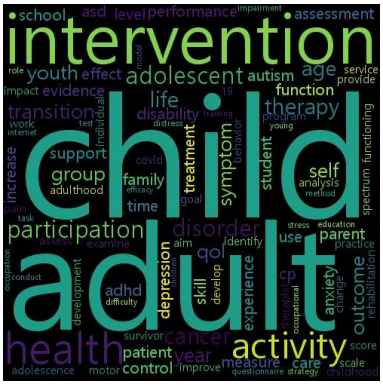

Figure 2-b. late adolescent TF- IDF

3.2 Analysis of co-occurrence word (Network analysis and centrality analysis)

After analysis of co-occurrence of 'late adolescent' word network, connection strength of 'child→ activity'(1340) was highest, and others followed by 'child→intervention(1314)', 'child→participation(1268)', 'age→year(1211)', 'intervention→health(952)'.Generally, research subjects related to activity and health of adolescents were recorded with the highest connection strength. Especially, multiple word connections with high connection strength were found such as 'child→activity', 'health→activity(901)', 'adult→health(883)' for word connections related to activity. This means that the studies related to adolescents' activity were actively conducted until now (Figure 3).

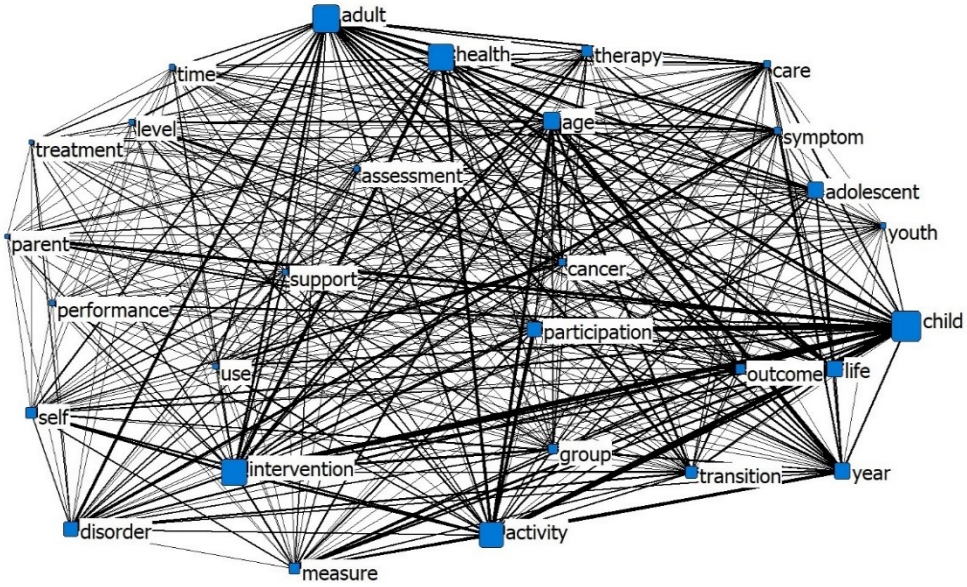

Figure 3. Network connection weight of late adolescent

3.3 LDA analysis

LDA based topic modeling method was used to deduce appropriate topic subject to classify text data. Topic modeling is the method to extract topic, which is to extract the core thesis from the text document group [19]. Statistical calculations were made to find out distribution of words in corpus, which is composed of documents. From the distributions, multiple topics were extracted and, in each topic, main words were allocated to distinct them by meaning. To deduce the optimal count of topics, coherence value and perplexity value, which are the scale to determine the abundance of similar words in one topic, were used to deduce the optimal count of topics. [20]. The higher the coherence value, the consistency of meaning between words in the subject. Meanwhile, the lower the perplexity value, the more the topic model reflects the document content. However, the analysis interpretation is not at ease even if the perplexity value is low, and generally the adequate amount of topic quantity is decided with the coherence value. Also, as the topic count increases, the content of the topic has a high change of reflecting the document content, but will be against the meaning of topic extraction [21]. Keywords in each topic are words that are allocated and related to each topic during topic modeling phase, but can repetitively appear in other topics. The meaning of extracted topics can be determined by key words, but it is not easy to classify into topic, understand the meaning, and naming as they are overlapped in other topics. Distinctive words are words that only appear in its topic and not in other topics, so it is effective to utilize distinctive words to proceed classification of meaning and naming implementation [22]. Therefore 8 topics were extracted after comprehensively considering the coherence and perplexity value and document content reflectivity, and meaning of topic extraction, and the topic name designation was classified by utilizing words related to topics and distinctive words. If you look at the topics, topic 1 includes ‘adolescent’, ‘health’, ‘child’, ‘life’, ‘activity’, etc., and different from other topics, ‘youth’, ‘task’, etc., appear as distinctive words, and was able to acknowledge the subject of the content as to be lifestyle of late adolescents.

Topic 2 includes ‘adult’, ‘ADHC’, ‘symptom’, ‘disorder’, etc. and distinctive words as ‘ADHC’, ‘disorder’, so it showed it was related to treatment psychiatric symptoms such as ADHC. Topic 3 includes ‘activity’, ‘adult’, ‘participation’, ‘child’, ‘life’, etc., and as for ‘participation’, ‘therapy’, it was shown that it was related to specific activity participation leading to treatment. Topic 4 includes ‘health’, ‘year’, ‘age’, ‘adolescent’, etc. and ‘year’, ‘care’ was shown so that the subject was related to treatment related to the age of late adolescents. Topic 5 shows ‘child’, ‘intervention’, ‘activity’, ‘ASD’, and distinctive words as ‘intervention’, ‘ASD’, which showed that the topic was related to disorder, and intervention.

4. DISCUSSION

This study seeks to understand the study trends of activity participation, treatment, and mediation of late adolescents through big data analysis methods. First, TF analysis result shows, ‘child(729)’ ranked highest, and followed by ‘adult(672)’, ‘intervention(651)’, ‘health(640)’, ‘activity(618)’.The words related to period such as ‘transition’, ‘time’ was also derived, and this was due to studies conducted for period adolescents turned to adults. Besides this, words related to self-management of adolescents such as ‘self(347)’, etc were also derived. Generally activity, treatment, parental support of adolescents related words were derived and the studies were actively conducted on them. [23]. TF-IDF analysis result shows, ‘child(25.59666)’, raked highest, followed by ‘adult(22.45008)’, ‘intervention(22.12035)’, ‘health(21.06596)’, ‘activity(20.02625)’. Words related to treatment such as ‘cancer(15.83276)’, ‘therapy(15.16334)’, ‘symptom(11.52597)’, care(10.93194)’, were also derived in top 10 to 30 list. Besides this, parent related words such as ‘parent(10.13155)’, were derived. TF and TF-IDF results both show that generally research subject words related to activity, treatment, and parents of adolescents were all ranked high. Second, co-occurrence word network analysis result shows

connection strength was highest for 'child→activity(1340)', and was followed by 'child→intervention(1314)', 'child→participation(1268)', 'age→year(1211)', 'intervention→health(952)'.In network analysis, the centrality of connection strength is high when the links connected to nodes were high. This was used as specific node centrality metric rather than the total network centrality [24]. The centrality of connection strength of late adolescents result shows that, 'child' was recorded highest (255.702), and followed by 'adult(230.053)', 'intervention'(236.86), 'health'(200.456), 'activity'(223.018). Thus, adult has the most connectivity compared to other words, and the connected words were 'health', 'therapy', 'disorder', 'intervention', which shows that research were actively conducted on discipline, regulations, health, and treatment [25]. Also, high in centrality of words were 'intervention(236.86)', 'health(200.456)', activity(223.018)' Third Appropriate topic subject was derived by using LDA(Latent Dirichlet Allocation) based topic modeling to classify text data. For topic 3 which was related to specific activity participation topic 3 [26], the highest topic ratio was 31.5%. Topic 5 related to intervention and disorder, the highest was 22.5%. Topic 4 related to age of late adolescents, the highest was 19.9%. Topics generally included activity related words such as 'activity', etc, and this showed that treatment, lifestyle, and age related to activity of adolescents were utilized as main research subject.

5. CONCLUSION

This study seeks to determine the research trend of late adolescents by utilizing big data. Also, seek for research trends related to activity participation, treatment, and mediation to provide academic implications. For this process, gathered 1.000 academic papers and used TF-IDF analysis method, and the topic modeling based on co-occurrence word network analysis method LDA (Latent Dirichlet Allocation) to analyze. The conclusion for this study is as below. The visualization was conducted with word cloud method by analyzing the morphemes with the gathered data, and TF analysis result shows 'child(729)' recording highest, and followed by 'adult(672)', 'intervention(651)', 'health(640)', 'activity(618)'. TF-IDF analysis result showed that 'child(25.59666)' ranked highest, and followed by 'adult(22.45008)', 'interventiona(22.12035)', 'health(21.06596)', 'activity(20.02625)'. The co-occurrence word network analysis result showed, connection strength of 'child→ activity' ranked highest, and result of topic modeling based on LDA(Latent Dirichlet Allocation) showed that it was classified into 5 keywords such as late adolescent identity, late adolescent symptom, late adolescent intervention, and etc. In conclusion this study conducted analysis of activity participation, treatment, and mediation of late adolescents by TF-IDF analysis method, co-occurrence word network analysis method, and topic modeling analysis based on LDA(Latent Dirichlet Allocation). The results were proposed through visualization, and carries significance as this study analyzed activity, treatment, mediation factors of late adolescents, and provides new analysis methods to figure out the basic materials of activity participation trends, treatment, and mediation of late adolescents.

ACKNOWLEDGEMENT

This paper was supported(in part) by Research Funds of Kwangju Women's University in 2022 (University Innovation Support Project.

REFERENCES

[1] J. Arnett, "Emerging Adulthood: A theory of development from the late teens the teens. *American Psychologist*, 55, 469-480, 2000.
[2] J. Arnett, "Emerging Adulthood." *Chronicle of Higher Education* 51(12): 1-8, 2004.
[3] M. Law, Participation in the occupations of everyday life". *American Journal of Occupational Therapy*,

56, 640–649, 2002.
[4] World Health Organization. *International classification of functioning, disability, and health*. Geneva, Switzerland: WHO, 2001.
[5] C. M. Baum, and M. Law, "Occupational therapy practice: Focusing occupational performance". *American Journal of Occupational Therapy*, 51, 277-288, 1997.
[6] A. A Wilcock, "Relationship of occupations to health and well-being". *Occupational therapy: Performance, participation, and well-being* (3rd ed.), 2005.
[7] M. Law, "Participation in the occupations of everyday life". *American Journal of Occupational Therapy*, 56, 640–649, 2002.
[8] N. W. Kim, & H. R. Choi, "Analysis of trends in domestic governance research through social network analysis", *Paper of the Korean Digital Policy Society, 16(*7): 35-45, 2018.
[9] C. M. Baum, & C. H. Christiansen, "Person-environment occupational- performance: An occupation-based framework for practice". 3-267), Thorofare, NJ: Slack, 2005.
[10] S. E. Lee," Study of Korean Modern Cookbooks Using Text Mining Analysis", 2022.
[11] A. Karl, J. Winooski, W. H. and Rushing, "A practical guide to text mining with topic extraction". *Wiley Interdisciplinary Reviews: Computational Statistics,* 7(5): 326-340, 2015.
[12] M, OH. Jang, S and UM. Kim, "Article analytic and summarizing algorithm by facilitating TF-IDF based on k-means". *Proceedings of the Korea Information Processing Society Conference,* (0): 271–274, 2018.
[13] J. Y. Lee, "Analysis of dance performances in the Covid-19 using big data on social media" Department of performing Arts and Multimedia Graduate School, Kookmin University, Seoul, Korea, 2021.
[14] Oh. Min, J. O. Kim, Y.G. Kim et al., "A Comparative Analysis for the Extraction of Similar Patent Claims based on Word Embedding", *Proceedings of The Korean Institute of Information Scientists and Engineers*, 100-1003, 2017.
[15] B. M. Kang, "Constructing Networks of Related Concepts Based on Co-occurring Nouns". *Korean Semantics,* 32(2): 1-28, 2010.
[16] M. L. Doerfel, "What constitutes semantic network analysis?" *A comparison of research and methodologies. Connections*, 21(2),16-26, 1998.
[17] J. H. Hong, and E. K, Na, "Victim Blaming of Sewal-ferry Disaster on News in Conservative Total TV Programming - Categorization of Victims and Word Network Analysis", *Korean Society for Journalism and Communication Studies,* 59(6), 69-106, 2015.
[18] J. S. Park, S. G. Hong, and J.W. Kim, "A Study on Science Technology Trend and Prediction Using Topic Modeling", *Journal of the Korea Industrial Information Systems Research,* 22(4):19-28, 2017.
[19] T. Hofmann, "Probabilistic Latent Semantic Analysis", *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc*, pp.289-296,1999.
[20] Y. S. Hun, K. H. Kim, "Expansion of Topic Modeling with Word 2 Vec and Case Analysis". *The Journal of Information Systems*, 30(1), 45-64, 2021.
[21] C. W. Jeong, and J. J. Kim, "Analysis of trend in construction using text mining method". *Journal of the Korean Digital Architecture and Interior Association*, 12(2), 53-60, 2012.
[22] H. J. Jung, "Research dynamics in innovation studies using text mining". *Journal of Technology Innovation*, 24 (4), 249-276. doi:10.14383/SIME.2016.24.4.249, 2016.
[23] K. Bowman, Development of Activity Card Sort for children: *Does parental report on the Activity Card Sort reflect similar results of their children.* Unpublished master's degree, University of Western Ontario, London, 1999.
[24] D. J. Kang, and K. N. Lee, "A Study on Co-author Networks of Journal of Korea Trade Research Association using Social Network Analysis". *Korea Trade Research Association,* 40(5): 1-23, 2015.
[25] H. Gansu, "Children's engagement in the world". *Sociocultural perspectives* pp. 148-170, 1999.
[26] C. L. Hilton, M. C. Crouch, H. Israel, et al, "Out-of-school participation patterns in children with high-functioning autism spectrum disorders". *American Journal of Occupational Therapy, 62,* 554-563, 2008.