New Physics: Sae Mulli,
Vol. 67, No. 5, May 2017, pp. 555∼561
http://dx.doi.org/10.3938/NPSM.67.555

Trend Analysis by Using Text Mining of Journal Articles Regarding Consumer Policy

Min-Jeong Kim · Kyungyoung Ohk* · Chung-Sook Moon
Department of Consumer Economics, Sookmyung Women's University, Seoul 04310, Korea

(Received 10 January 2017 : revised 3 April 2017 : accepted 4 April 2017)

This paper identifies the direction of change in consumer policy research, where consumer policies are shifting from consumer protection perspectives to consumer responsibility perspectives. In this research, text mining and association-rule techniques were used on the titles and keywords of articles published in "Consumer Policy and Education Review", a major journal for consumer policy studies, from 2005 to 2015 to find major research trends, longitudinal trends, and a research field association rule; these results were visualized so that they could be easily understood. First, 1026 keywords were extracted from 238 papers that had titles and keywords. From these, 73 keywords that appeared 10 times or more were selected as research targets and were analyzed. From the analysis, the longitudinal patterns of the keywords' appearance frequency by year showed steady increasing, increasing, one-time peak, and rapidly increasing then decreasing trends. Finally, the association rule was used to find frequent patterns in the keyword database; this revealed a structure in which groups were formed based on the keyword "consumer". Through this research, the trends and the keywords of papers on consumer policy studies could be understood; thus, this research can be used as fundamental data for predicting trends in articles on consumer policy studies.

PACS numbers: 89.70.+c, 89.90.+n1
Keywords: Consumer policy's trend, Text mining, K-means clustering, Association rule

I. INTRODUCTION

Text mining refers to a process of extracting meaningful, non-trivial patterns or knowledge from a set of unstructured texts [1]. Identification of meaningful patterns and trends and the extraction of potential knowledge in large volumes of text data is an important task in various fields [2]. In particular, the advent of high-speed internet generates large amounts of textual data in a variety of forms [3]. As an aspect of this trend, in various fields such as academic article information and news article information, research utilizing text mining technique is actively being carried out to find trends and extract implicit information from large volume of data [1,4,5].

This paper identifies the direction of change in consumer policy research, where consumer policies are shifting from consumer protection perspectives to consumer responsibility perspectives. In this research, we analyze "Consumer Policy and Education Review" journal during 11 years from its first issue (2005) to determine its main concerns and research concentrations, the trends that have appeared in its research over the past 11 years, how they have changed, and which research subjects are connected. This research uses text mining and association rule techniques to analyze trends in articles published in "Consumer Policy and Education Review".

Through this, the major research trends and longitudinal trends, and keyword co-occurrence information were determined and visualized to promote ease of understanding. As such, this research has been able to observe the status and progress of research in consumer policy

*E-mail: okyoung@sookmyung.ac.kr

Table 1. Analysis target.

Year	Published Articles	Number of Issues
2005	8	1
2006	15	2
2007	13	2
2008	24	4
2009	21	4
2010	21	4
2011	28	4
2012	20	4
2013	27	4
2014	35	4
2015	26	3*
Total	238	

*Analysis target of year 2015 include 3 issues

across time, and it has been able to examine the interrelations among the keywords included in "Consumer Policy and Education Review".

II. DATA GATHERING

"Consumer Policy and Education Review" was first published in 2005. In 2005, it was released in a volume, number format one time. Since 2007, it has been released in the volume, number format quarterly. The analysis target articles are outlined in Table 1. We used the titles and keywords of the 238 articles as targets for text mining analysis. In this research, the publication year, article title, and article keywords are used as the main data for performing an analysis of keywords by article and an analysis of keywords by year.

III. ANALYSIS METHODS

The program used for text mining and association rule was an open source program called R, version 3.2.3. The tm (text mining) package [6] was used to perform text mining on the titles and keywords of the previously collected articles. Based on this, a term-document matrix and a term-year matrix were created. Then, k-means clustering [7], a nonhierarchical clustering method, was used to understand how keywords have changed, and which research subjects are connected. The arules (association rules) package [8] was used to find relations among keywords, and the arulesViz package [9] was used to visualize the correlations.

The R program's tm package can analyze a data structure called a corpus, so the txt data files were read and converted into a corpus, and then basic text processing and subsequent analysis were performed. The basic text processing converts all text into lowercase, removes symbols such as numbers and punctuation, removes stop words, and removes blank space. After this, it is possible to construct a term-document matrix that describes the frequency of keywords that occur in a collection of documents. Then, we selected important keywords that appeared frequently, excluding words with multiple meanings, common verbs, and general words. Based on these important keywords selected, the high dimensional datasets were created from the term-document matrix made with the text mining results. The other dataset is the portion of data representing the frequency of the keywords over time period from 2005 to 2015 and called term-year matrix.

Generally, the term-document matrix simply uses the frequency by keyword, but in this research, rather than using a simple document appearance frequency by keyword, a weight was calculated for each word with regard to Term Frequency-Inverse Document Frequency (TF-IDF), which is widely used in information search and text mining research [10]. This is a value that can judge how important a particular term is in a particular document from a collection of several documents. Term Frequency (TF) is a value representing how frequently a particular word appears in a document; the larger this value, the more important a word can be considered to be in a document. Frequent use in a document collection means that a word appears commonly. Therefore, not only the word frequency but also the document frequency is considered. This is called Document Frequency (DF), and the reciprocal number of this value is called Inverse Document Frequency (IDF), which has a weight for each word. IDF is a value that shows how commonly a certain word appears in a document collection. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term,

and then taking the logarithm of that quotient. Therefore, TF-IDF is a value that multiplies TF and IDF and is calculated as follows.

$$\text{TF-IDF} = \text{TF} \times \log(\text{N}/\text{DF}) \quad (1)$$

- TF : Frequency of a particular word in the documents
- N : Number of all documents
- DF : Number of documents containing a particular word
- IDF : Inverse of DF

The IDF value increases if few documents contain the specified word among all documents, and it decreases if many documents contain the specified word among all documents. This is because as the number of documents that contain the word increases, the value of N/DF approaches 1, and the IDF value approaches 0.

Next, for the term-document matrix, k-means clustering was used to determine how clustering occurred in relation to specific keywords. Then, k-means clustering was also used to understand the yearly changes in specific keywords for the term-year matrix.

Finally, using association rule technique, we found the relations between keywords, where each document is considered as a transaction, and each keyword in the document is an item. There are three main measures called support, confidence, and lift [11] to assess the strength of the rule among specific keywords.

IV. ANALYSIS RESULTS

1. Main Keywords

A total of 1736 words were used in the titles and keywords of the 238 articles, but after the words not needed for analysis were removed, 1026 words remained. Looking at this process step-by-step, a corpus was first created using the data for analysis, then text processing was performed (removing blank space, punctuation, symbols, numbers and stop words; performing lowercase conversion; *etc.*) so that only the analysis target keywords were extracted. In the next step, the analysis target keywords were analyzed, and because they could be singular or plural, all plural words in the corpus were converted into singular words. Third-person verbs and past-tense verbs were converted into present-tense verbs, and other additional modifications were made to the corpus.

Table 2. Summary of articles and keywords from the "Consumer Policy and Education Review".

Category	Collection Targets	Research Targets
Articles	238	238
Keywords	1736	1026

Through this process, the 1736 words from the 238 articles were condensed into 1026 words; the collected items are listed in Table 2. In this research, the frequency analysis was based on the final 1026 keywords.

In this paper, we selected 73 important keywords that appeared more than 10 times, excluding words with multiple meanings, common verbs, and general words from the 1026 words that were gathered. For example, words such as "study", "analysis", and "research" were included in the most frequent keywords, but they had no value for analysis and thus were not chosen. The 73 selected keywords are listed in Table 3 in order of frequency. Using only TF value, we can understand which keywords are frequently used in the "Consumer Policy and Education Review" journal. The word "consumer" was overwhelmingly frequent in Table 3, because the "Consumer Policy and Education Review" journal is about consumer study. Looking at the keywords, there were many words related to consumer. Unlike other journals, the most frequent keywords in "Consumer Policy and Education Review" included very little on research methodology, and most words were on research subjects related to consumer's studies.

2. Main Keywords by Year

In this section, using tm package, we selected important keywords that appeared frequently by year. The selected keywords by year are listed in Table 4. Looking at the keywords, there were many words related to

Table 3. Most frequent keywords in the "Consumer Policy and Education Review" and their frequency.

Keyword	Fr.	Keyword	Fr.	Keyword	Fr.	Keyword	Fr.	Keyword	Fr.
consumer	420	Korea	30	plan	20	internet	15	perception	12
consumption	99	attitude	28	effect	19	use	15	recall	12
behavior	84	program	28	regulation	19	awareness	14	search	12
information	79	green	26	management	18	labeling	14	demand	11
education	77	knowledge	26	risk	18	protection	14	energy	11
financial	62	student	23	social	18	school	14	evaluation	11
satisfaction	53	system	23	development	17	case	13	literacy	11
service	41	advertising	21	ethical	17	comparison	13	CCM	11
food	40	personal	21	type	17	content	13	need	11
counsel	33	purchase	21	adolescent	16	act	12	stress	11
product	33	retirement	21	factor	16	certification	12	income	10
safety	33	value	21	local	16	child	12	north	10
focus	32	credit	20	mobile	16	debt	12	wellbeing	10
policy	32	customer	20	model	16	experience	12		
center	31	household	20	competency	15	life	12		

Table 4. Keywords by year.

Year	Keywords
'05	policy, information, act, protection
'06	education, personal, curriculum, debt, financial
'07	education, school, north
'08	education, behavior, satisfaction, financial, administration, retirement
'09	energy, advertising, information, internet, labeling, wellbeing
'10	green, food, counselling, network, safety, act, financial, knowledge
'11	safety, green, recall, education, attitude
'12	financial, retirement, food, information, policy, safety, green, knowledge
'13	credit card, eco-friendly, literacy, social, usage, commerce, risk, competency
'14	mobile, service, ethical, customer, management, CCM, certification, counselor
'15	regulation, annuity, boycott

consumer's right such as consumer protection and consumer education in early days, gradually the articles addressed consumer responsibility-related research, and the frequent keywords included "energy", "green", "eco-friendly", and "ethical". These results show that consumer policy research with regard to the trend of research has been changed from consumer protection perspectives to consumer responsibility perspectives.

3. Keyword Clustering

To examine the group patterns of the term-document matrix and the term-year matrix calculated in TF-IDF value, they were divided into clusters using k-means clustering. Table 5 shows how the clusters for certain keywords were divided using the term-document matrix. The numbers of groups (k) to be divided into were set at 3, 4, and 5. Division into 3 groups made it possible to obtain the most meaningful results.

In the clustering results of the term-document matrix, the categories appeared to be formed according to the keywords of the cluster. The keywords of Cluster 1 included a variety of keywords mostly about research related to general consumer. Cluster 2 was made of only four keywords, and all keywords ("food", "policy", "recall", and "safety") were related to the food safety. Cluster 3 also includes a variety of keywords about research related to financial education for a variety of target. Through this, it was discovered that the most frequent keywords in the "Consumer Policy and Education Review" journal could be broadly categorized as policies about general consumption and food safety, and financial education for a variety of target.

The "Consumer Policy and Education Review" journal published 36 issues from Volume 1, Number 1 in November 2005 to Volume 11, Number 3 in September 2015. Therefore, in this research, a term-year matrix was used

Table 5. K-means clustering results for term-document matrix.

Keywords	Category
Act, adolescent, advertising, attitude, awareness, ase, CCM, center, certification, child, comparison, competency, consumer, consumption, content, credit, customer, debt, demand, development, effect, energy, ethical, evaluation, experience, factor, focus, green, household, income, information, internet, Korea, labeling, life, local, management, mobile, model, need, north, perception, product, program, purchase, regulation, search, service, social, system, type, use, value, wellbeing	General consumer
Food, policy, recall, safety	Food safety
Behavior, counsel, education, financial, knowledge, literacy, personal, plan, protection, retirement, risk, satisfaction, school, stress, student	Financial Education

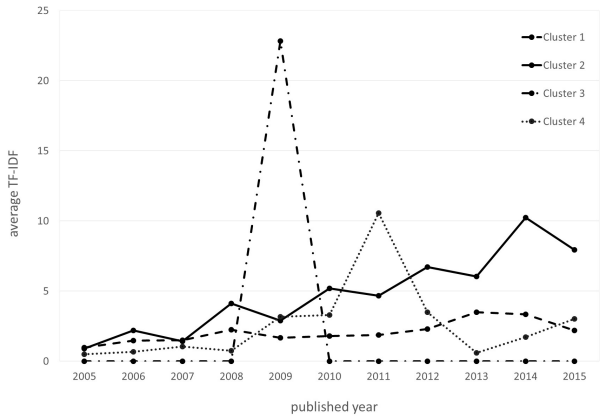

Fig. 1. Average TF-IDF of four clusters of the keywords observed over the past 11 years (2005-2015).

to find the occurrence patterns for each keyword from 2005 until 2015. k-means clustering was performed to investigate the changes in keywords by year. Fig. 1 shows the average TF-IDF factor of four clusters from k-means clustering.

Cluster 1 shows a trend of steady increase. Cluster 2 shows a trend of continual increase since 2005. Cluster 3 has one-time peak with the highest average TF-IDF factor in 2009, and it does not exist before and after 2009. This is believed to be because there was a special edition in 2009 on the topic of energy. Cluster 4 shows a trend of increasing rapidly in 2011, followed by a decrease.

A list of the keywords that belongs to each cluster is shown in Table 6. This result can provide an understanding of research trends in the "Consumer Policy and Education Review" journal over the past 11 years. Names were assigned to each of the cluster categories in Table 6 according to the yearly trends of average TF-IDF results for each of the clusters. Cluster 1 is the steady increasing keywords, Cluster 2 is the increasing keywords, Cluster 3 is the one-time peak keywords, and Cluster 4 is the rapidly increasing then decreasing keywords.

Looking at the keywords in Cluster 1, it can be seen that the keywords are distributed throughout a variety of research topics that are addressed in research related to consumer. The keywords that have been steadily receiving attention from researchers since 2005 are in Cluster 1. Looking at the keywords in Cluster 2 ("behavior", "consumer", "consumption", "education", "ethical", "financial", "food", "information", "mobile", *etc.*), it can be seen that the keywords receive a fairly large amount of attention. Cluster 3 showed the one-time peak in 2009 but has no more attention, which is believed to be because energy was the subject of a special edition in 2009. Cluster 4 showed a rapid increase in 2011 but has gradually lost attention, which has a distribution of keywords about "green", "safety", and "wellbeing".

4. Keyword Association Rules

This section analyzes the association rules for the 73 words that appeared in article titles and keywords. Table 7 shows the keyword support, confidence, and lift of the association rules between keywords. For this, rules were derived only with a support value of at least 0.01 and a confidence value of at least 0.5, but this produced a total of 963 association rules. Of these, 107 association rules with 1 element each were found. Of these, the 10 rules with the highest lift value are shown in Table 7.

We can know that most of these are made of words that are used associatively. For example, the support value for "debt" and "household" (*i.e.*, the probability that they will appear together) is small at 0.0168, but in articles that include the antecedent subject "debt," the

Table 6. K-means clustering results for term-year matrix.

Keywords	Category
act, case, CCM, certification, child, comparison, competency, content, credit, customer, debt, development, effect, evaluation, experience, factor, focus, household, income, internet, knowledge, Korea, labeling, life, literacy, local, management, need, north, perception, personal, plan, policy, program, protection, regulation, retirement, risk, school, search, social, stress, student, system, type, use	Steady increasing
behavior, center, consumer, consumption, education, ethical, financial, food, information, mobile, satisfaction, service	Increasing
Energy	One-time peak
adolescent, advertising, attitude, awareness, counsel, demand, green, model, product, purchase, recall, safety, value, wellbeing	Rapidly increasing then decreasing

Table 7. Keyword Support, Confidence and Lift.

Antecedent	Consequent	Supp.	Conf.	Lift
debt	household	0.0168	0.8000	19.0400
CCM	case	0.0252	0.6667	15.8667
Food	safety	0.0336	0.6667	14.4242
income	household	0.0168	0.5714	13.6000
north	Korea	0.0168	1.0000	12.5263
CCM	center	0.0336	0.8889	11.7531
case	customer	0.0210	0.5000	10.8182
certification	plan	0.0126	0.5000	9.9167
retirement	plan	0.0168	0.5000	9.9167
labeling	system	0.0168	0.5714	9.0667

confidence value of the consequent subject, "household," (*i.e.*, the probability that it will appear) is very high at 0.8. Furthermore, in the case of the lift—*i.e.*, the association—"debt" and "household" have a lift value of 19.04, so it can be known that if "debt" appears, the possibility that "household" appears is fairly great.

Fig. 2 shows the results of using a network graph to visualize the association rules between keywords. A variety of keywords are evenly associated with "consumer" at the center. Most of keywords are directly connected to "consumer" except for isolated keywords such as "debt",

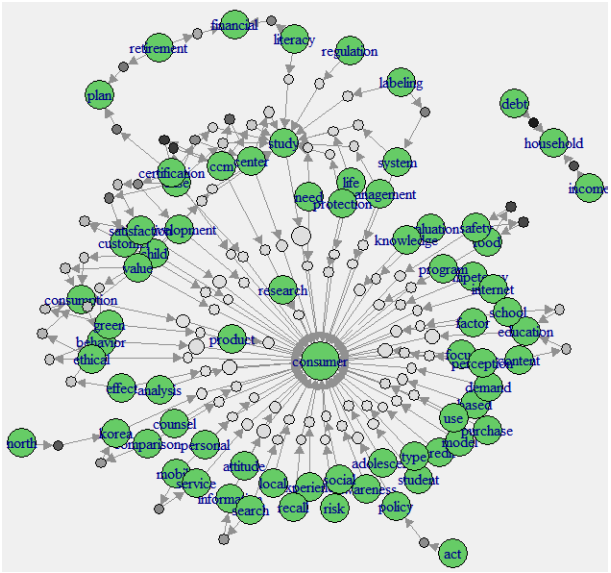

Fig. 2. (Color online) Keyword correlation analysis results.

"household", and "income". However, aside from these, there are other keywords ("north", "act", "labeling", and "regulation") that are indirectly connected to "consumer" through other keywords. Through this, it can be known that the "Consumer Policy and Education Review" journal deals with a variety of research topics centered on "consumer" with only a few isolated keywords. Moreover, the sizes of the circles in Fig. 2 are proportional to the support values between the two keywords, and the darker the circle's color, the larger the lift value. Therefore, it is easy to see whether there is a stronger association rule between keywords. It can be seen that the rule debthousehold has a highest lift value.

V. CONCLUSION

In this research, we analyzed the titles and keywords of 238 articles from the "Consumer Policy and Education Review" journal in order to determine its main concerns and research concentrations, the trends that have appeared in its research over the past 11 years, how they have changed, and which research subjects are connected. We utilized text mining and association rule techniques to analyze trends in articles published in "Consumer Policy and Education Review".

First, we found the 73 keywords that appear most frequently in the titles and keywords of the 238 articles,

and also selected main keywords by year. Then, k-means clustering was used on the term-document matrix to see how the clusters would form. Next, the results of clustering analysis based on a term-year matrix showed that there were keywords decreasing except for a particular year and keywords with one-time peak, but for all other keywords, the number of articles increased along a time series. Finally, the association rules between keywords were analyzed to quantify and visualize the relations of each keyword. The association rules results showed that a large group formed based on the keyword "consumer".

This research is significant in that it is the first to apply text mining and association rule mining techniques to consumer policy-related research articles. We could identify the direction of change in consumer policy research similar to consumer policy paradigm which is shifting from consumer protection perspectives to consumer responsibility perspectives. Moreover, the results of this research can be used before starting new research to find articles with keywords similar to those in existing research and seen which research topics the "Consumer Policy and Education Review" journal has been covering recently.

However, in this paper, text mining was performed on only titles and keywords. For a more comprehensive analysis, the abstracts of articles need to be included. Unfortunately, the data of this journal were collected manually because data from "Consumer Policy and Education Review" journal website could not be downloaded as html files. For future follow-up research, it is necessary to perform analysis with more data including abstracts and comparative analysis on other data such as news articles related to consumer policy to determine their similarities and differences.

REFERENCES

[1] J. L. Hung and K. Zhang, J. Comput. High. Educ. **24**, 1 (2012).

[2] A. Kao and S. R. Poteet, *Natural Language Processing and Text Mining* (Springer-Verlag London, London, 2007).

[3] S. G. Cho and S. B. Kim, Int. J. Inf. Educ. Technol. **2**, 233 (2012).

[4] S. J. Lee, S. H, Lee, H. J. Seol and Y. T. Park, R&D Management **38**, 169 (2008).

[5] A. Balahur and R. Steinberger, in *Proceedings of the 1st Workshop on Opinion Mining and Sentiment Analysis* (Seville, 2009).

[6] Package 'tm', https://cran.r-project.org/web/ packages/tm/tm.pdf (accessed Jan. 10, 2016).

[7] Package 'cluster', https://cran.r-project.org/web/ packages/cluster/cluster.pdf (accessed Jan. 10, 20 16).

[8] Package 'arules', https://cran.r-project.org/web/ packages/arules/arules.pdf (accessed Jan. 13, 20 16).

[9] Package 'arulesViz', https://cran.r-project.org/we b/packages/arulesViz/arulesViz.pdf (accessed Jan. 13, 2016).

[10] G. Salton and C. Buckley, Information Processing and Management **24**, 513 (1988).

[11] R. Agrawal, T. Imieliński and A. Swami, in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (Washington, D.C., 1993).