Indian Journal of Science and Technology, Vol 9(25), DOI: 10.17485/ijst/2016/v9i25/97151, July 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

An Analysis of Radiation Fusion Technology-Related Patents Using Statistical Methods and Data Mining Techniques

Seung-Joo Lee[1], Tae-Jong Han[2] and Kyoungho Choi[3*]

[1]Department of Statistics, Cheongju University, Korea; access@cju.ac.kr
[2]Department of Radiological Science, Jeonju University, Korea; tjhan@jj.ac.kr
[3]Department of Basic Medical Science, Jeonju University, Korea; ckh414@jj.ac.kr

Abstract

Background/Objectives: This study was conducted to identify the trends related to the patents for 'radiation anti-oxidation technology', and to perform technological forecasting for that technology using statistical methods and data mining techniques. **Methods/Statistical Analysis:** The extraction of patents for analysis was carried out using a website KIPRIS. Documents including the words 'irradiation' and 'anti-oxidation' were searched for to target patents registered between January 1999 and January 2015. Finally, a total of 512 patent documents were selected as analysis objects through an editing process. **Findings:** Key finds are as follows. First, most of the top 10 patents are related to cosmetic development, drug development and food processing. Second, in terms of the degree of support, the degree is high when a technology of A61Q 19/08 is first developed and then a technology of A61K 8/97 is developed. Third, in the top association rules identified using 491 IPC codes included in 512 RFT-related patent documents, the technologies of "proliferation of flowering plants using tissue culture technology" are developed. **Application/Improvements:** The findings of this study are significant as they have derived basic data for technological forecasting by identifying specific information about core technology factors included in each patent for radiation anti-oxidation technology.

Keywords: Data Mining, IPC Code, Patent, Radiation Fusion Technology, Statistical Analysis

1. Introduction

Radiation is one of the myriad elements and constituents that makes up the natural world, like water and air, and it exists everywhere in the world. Human beings make radiation artificially and use radiation in a wide range of areas such as the diagnosis and treatment of diseases, to improve the quality of farm produce, for disinfection and sterilization purposes, food preservation, and the non destructive testing of industrial products. Recently, as the applications of radiation have expanded, RFT is being actively used in combination with other academic areas.

RFT, which is a cutting-edge radiation fusion technology that involves the fusion of radiation with IT, BT, NT and ET, creates new added value by developing radioactive new medicine or radiation healthcare technologies and eco-friendly radiation fusion technologies. In particular, the development of new and high value-added bio-products using the convergence of radiation technology with BT and the development of medical and industrial technologies using bio-material production technology and radiation genomics are emerging as a core technology of the 21st century bio-tech industry. Thus, as RFT becomes more and more important, it is very meaningful to forecast new technologies by systematically analyzing the growing amount of relevant information.

The existing methods of technological forecasting are classified into two main categories[1]. The first is a qualitative analysis method such as the tree method[2] and the Delphi method[3] that identifies technology trends through the meeting and coordination of views among experts. The second is a quantitative analysis method such as trend analysis[4] and patent analysis[5-7]. Among these, patent

*Author for correspondence

analysis is an approach for deriving information useful for a specific technology by collecting and analyzing patent documents including the results of the development of the technology and to effectively carry out technological forecasting using the results of the patent analysis. Specific methods of patent analysis include a method based on descriptive statistics such as frequency analysis, citation analysis and visualization and a method based on inductive statistics such as time series analysis, regression analysis and Bayesian network model. Studies are recently being conducted on the processes of developing new technologies to develop new products through the application of principal component analysis and text mining techniques to keyword data included in patent documents[8].

As today's development of scientific technique and information technology has brought about changes in economic paradigms such as the shift of social and economic business areas from the industrial economy to the knowledge-based economy, new value creation through continuous innovation based on technological forecasting, the strengthening of core competence, the development of core technologies, and technological convergence is becoming more important[9].

In a move to apply this current trend to the field of radiation, this study intends to identify the trends related to the patents for 'radiation anti-oxidation technology', which is a type of RTF, and to perform technological forecasting accordingly. To this end, this study intends to collect the related patent documents and to conduct association rule mining and social network analysis using statistical methods such as principal component analysis and cluster analysis, and the text mining technique that is among data mining techniques. Further, this study intends to identify specific information about core technology factors included in each patent for radiation anti-oxidation technology and also to identify the relationships among technologies, central technologies and patents thereof. Accordingly, it is thought that the findings of this study can be used as basic data for technological forecasting.

2. Research Method

2.1 Data Collection

This study was conducted to analyze domestic patents among patents related to anti-oxidation technology included within fusion technologies using radiation. The extraction of patents for analysis was carried out using an website Korea Intellectual Property Information Service (KIPRIS) (http://kipris.or.kr). Documents including the words 'irradiation' and 'anti-oxidation' were searched for to target patents registered between January 1999 and January 2015. Finally, a total of 512 patent documents were selected as analysis objects through an editing process. KIPRIS is a domestic patent search web site, which enables the free searching of information through the data-basing of the information on industrial property right (e.g. rights of trademark, utility, design and patent) by Korea Institute of Patent Information.

2.2 Research Problem

This study intends to conduct an analysis on the following specific contents, using 512 patents selected as analysis objects.

First, this study intends to conduct patent analysis using the overall structure of IPC code ranging from sections to sub-groups in the preprocessing of IPC code. Accordingly, the data matrix of patent-IPC code is created through the preprocessing of the preceding patents used by each patent up to the sub-groups of IPC code. Also, a data matrix of patent-words is created through the preprocessing of English titles of patents.

Second, the trends related to patent registration and IPC code is first reviewed using patent documents and the preprocessed data matrix of patent-IPC code.

Third, patent documents is clustered using the data matrix of patent-words.

Fourth, patents with high degrees of support, confidence and lift is identified and technological forecasting is made through technology association analysis conducted using association rule mining.

2.3 Analysis Tool

R 3.1.2 was used in the processes of preprocessing and data mining for analysis including the arrangement of collected data[10,11].

2.4 Preprocessing

Some patent documents should be preprocessed before analyzing the patents for irradiation and anti-oxidation technologies using the 512 patent documents selected as analysis objects.

First, to analyze key words, it is necessary to create the data matrix of patent-words which is patent data structured from the 512 patent documents collected to analyze

the patents for radiation anti-oxidation technology in the first place. In this case, text mining techniques can be used as a preprocessing tool which can extract data from patent documents or convert patent documents into data so that patent documents can be applied to existing statistical analysis techniques. So, the data matrix of patent-words was created through the preprocessing of patent titles, which are the titles of inventions, from 512 patent documents collected for patent analysis using text mining techniques. Patent titles exist in the form of "Korean title (English title)" and Korean titles make it difficult to extract key words. Thus, key words were extracted from patent documents based on English titles. Each row of a created data matrix of patent-words indicates the patents for radiation anti-oxidation technologies and each column indicates extracted words. The data matrix of patent-words about the patents for radiation anti-oxidation technologies consists of 512 rows and 1,540 columns. The elements of the data matrix of patent-words indicate the frequency of words represented in each patent, and the number of words represented more than once in 512 patents is 1,540.

Second, this study intends to conduct patent analysis using a complete classification sign that represents the overall structure of IPC codes ranging from sections to sub-groups in the preprocessing of IPC codes. Accordingly, the data matrix of patent-IPC code was created through the preprocessing of the preceding patents used by each patent up to the sub-groups of IPC codes. Each row of a created data matrix of patent-IPC code indicates technology patents and each column indicates extracted IPC codes. The data matrix of patent-IPC code about the patents for radiation anti-oxidation technologies consists of 512 rows and 491(57) columns. The elements of the data matrix of patent-IPC code indicate the frequency of IPC codes represented in each patent, and the number of words represented more than once in 512 patents is 491(57).

3. Analysis Result

3.1 The Registration of RFT (Irradiation and Anti-Oxidation Technology) Patent and the Trend of IPC code

This study was conducted to analyze domestic analysis data about anti-oxidation among fusion technologies using radiation. Figure 1 shows the number of domestic

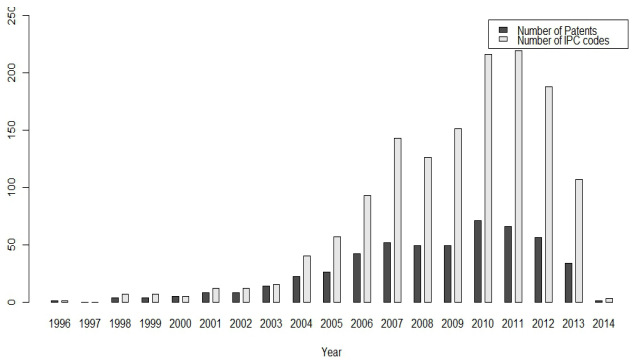

Figure 1. Number of patent registration and IPC code of irradiation and anti-oxidation technology.

patent registrations related to RTF over the last 20 years and the number of IPC codes. In the case of domestic patent registrations related to anti-oxidation technology among fusion technologies using radiation, the first patent was filed for in 1996. The number of patent registrations began to increase in early 2000 and continued on a steady upward trend until 2010 (71 patent registrations), while it began to decrease in 2011 (66 patent registrations) and was the lowest with only one in 2014. Patent-IPC code matrix created through preprocessing shows that the number of IPC codes by year was larger with 219 in 2011, 216 in 2010 and 188 in 2012 than the number of patent registrations.

3.2 Clustering Technologies using Keyword Analysis

This study intends to cluster the patents for radiation anti-oxidation technologies using a keyword analysis. Specifically, this study intends to cluster patents using the data matrix of patent-words created through the preprocessing of 512 patent documents collected to analyze the patents for radiation anti-oxidation technologies. The data matrix of patent-words concerning the patents for radiation anti-oxidation technologies consists of 512 rows and 1,540 columns. Each row of a created data matrix of patent-words indicates the patents for radiation anti-oxidation technologies and each column indicates extracted words. The elements of the data matrix of patent-words indicate the frequency of words represented in each patent, and the number of words represented more than once in 512 patents is 1,540.

A word cloud is a good tool to easily identify the important contents of patent documents visually by representative words with a high frequency among 1,540 words

used in patent documents. Figure 2 is a word cloud where words included in English titles of patent documents are visually represented.

The top 10 high frequency words used in the patent titles of RFT-related patent documents include composition, method, comprising, extract, thereof, containing, using, ginseng, preventing and pharmaceutical. RFT is chiefly used to develop the method of synthesizing products with medicine ingredients extracted from health foods such as ginseng.

Meanwhile, as the data matrix of patent-words has a great many words against the number of patents, a number of elements of the data matrix have a value of 0, which raises the problem of sparsity. In this case, it can cause many problems in directly using the data matrix of patent-words for statistical analysis. To solve these problems, a principal component analysis was conducted on the data matrix of patent-words, which was converted into the data matrix of patent-principal component scores along with dimension reduction. According to the results of the principal component analysis on the data matrix of patent-words, 331 principal components account for more than 95% of the total variations and the number of retained principal components is 331. The data matrix of patent-principal component scores was created using these principal components.

k-medoid clustering method, which is a non-hierarchical clustering method, was used as a method of clustering for patent analysis. To carry out k-medoid clustering, it is necessary to determine the optimal number of clusters of the data matrix of patent-principal component scores in the first place. Reference[12] proposed a silhouette statistic quantity to evaluate clusters and to determine the optimal number of clusters. Reference[13] proposed that a value to maximize the average silhouette width for the total data sets should be selected in determining the optimal number of clusters.

Accordingly, Figure 3 shows the results of calculating the average silhouette width along with the changes in the number of clusters (k = 2, 3, ..., 12) to determine the optimal number of clusters when k-medoid clustering is carried out using the data matrix of patent-principal component scores. As shown in Figure 3, when k, the number of clusters, was 2, the average silhouette width was the largest with 0.05. Therefore, the optimal number of clusters was determined as 2.

To cluster 512 patents as analysis objects, cluster analysis was conducted using a k-medoid algorithm and the data matrix of patent-principal component scores[14].

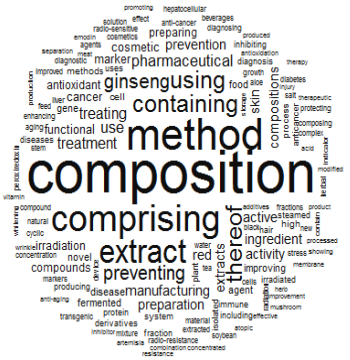

Figure 2. Word cloud of english titles of patent documents.

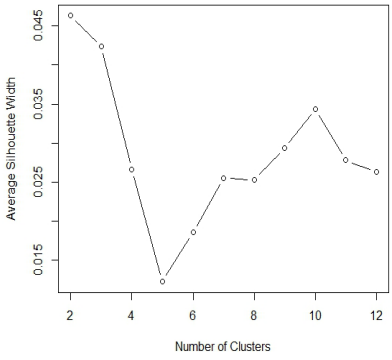

Figure 3. Average silhouette width.

Table 1. Cluster of RFT patents and patented technologies

Cluster	The Number of Patents	Percent-age	Top 10 Key words	Definition of Technology
1	324	63.3	method, using, ginseng, thereof, extract, use, red, manufacturing, compositions, containing	Technology about the method of extracting specific substances from ginseng, etc and of manufacturing synthetic material using those substances
2	188	36.7	composition, comprising, extract, containing, preventing, pharmaceutical, treating, active, method, ingredient	Technology about the activity or method of synthesizing materials with extracted medicine ingredients

Also, number of clusters was determined as 2. As a result of cluster analysis, 324 and 188 patents were allotted to cluster 1 and 2 respectively. Table 1 shows the results of defining each cluster using the top 10 key words extracted from patents pertaining to each cluster.

3.3 Technology Association Analysis

IPC codes were preprocessed up to sub-groups using 512 patent documents extracted for an analysis of RTF patents. Consequently, the data matrix of patent-IPC code was created with 512 rows and 491 columns, and a set of transaction data was created to be finally used. A total of 491 mutually different IPC codes included in the 512 patent documents become the components of a set of items for association rule mining. Table 2 shows the distribution of the number of IPC codes of each patent document: for the number of IPC codes per patent document, 77 cases on smallest size, 1 and 1 case on largest size 6; size 3 or 4 of IPC codes were concentrated in about 59% of the 512 patent documents, and their median was 3 and their mean was 2.734. It was found that 3 technologies were largely used in each patent document.

Figure 4 shows the top 10 high frequency IPC codes among 491 IPC codes included in patent documents.

Table 2. Distribution of the number of IPC codes of each patent document

Size	1	2	3	4	5	6	Total
The number of IPC codes	77	134	151	149	0	1	512

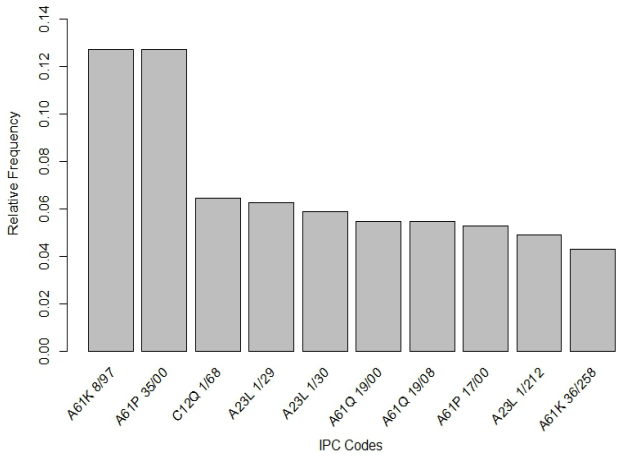

Figure 4. Top 10 high frequency IPC codes in RFT-related patents.

Table 3. Technologies of top 10 IPC code

IPC code	Frequency	Technology
A61K 8/97	65	Emersion of cosmetics or quasi-cosmetics derived from herbs (e.g. herbal extracts)
A61P 35/00	65	Antineoplastic agents
C12Q 1/68	33	Measurement or test methods using enzymes or microbes containing nucleic acids
A23L 1/29	32	Nutritional improvement of food: manufacturing and treatment of dietary food and groceries
A23L 1/30	30	Manufacturing and processing of food and groceries containing additives
A61Q 19/00	28	Skin care agents
A61Q 19/08	28	Antioxidant agents of Skin care agents
A61P 17/00	27	Dermatology drugs
A23L 1/212	25	Processing of beans for food production and processing of fruits or vegetables
A61K 36/258	22	Manufacturing of drugs with unknown structure containing ingredients derived from genus panax

Table 4. Top association rules using the degree of support

X→Y	rank	support	confidence	lift
A61Q 19/08→A61K 8/97 A61K 8/97 →A61Q 19/08	1	0.0508	0.9286 0.4000	7.3143
A61Q 19/00→A61K 8/97 A61K 8/97 →A61Q 19/00	2	0.0469	0.8571 0.3692	6.7516
G01N 33/574→C12Q 1/68 C12Q 1/68 →G01N 33/574	3	0.0293	0.8824 0.4545	13.6898

According to RFT-related patent documents, A61K 8/97 and A61P 35/00 are the fields of technologies that have been developed the most until now, followed by C12Q 1/68, A23L 1/29 and A23L 1/30. Table 3 defines the technologies pertaining to the top 10 IPC codes in the IPC 2015 version of Korean Intellectual Property Office (KIPO).

It can be seen that the technologies used the most in RFT are the technologies for "emersion of cosmetics or quasi-cosmetics derived from herbs (e.g. herbal extracts) (A61K 8/97)" and "antineoplastic agents (A61P 35/00)." This study used the APRIORI algorithm as an algorithm for association rule mining (Agrawal and Srikant,

1994). The number of association rules with the degree of support more than 0.01 and the degree of confidence more than 0.1 were 53. Among these association rules, the top 3 association rules based on the degree of support are shown in Table 4.

Although the degree of support of A61Q 19/08→A61K 8/97 is the same as that of A61K 8/97 →A61Q 19/08, the former is far larger in the degree of confidence than the latter. Therefore, the first significant technologies have a large degree of confidence when a technology of A61Q 19/08 is first developed and then a technology of A61K 8/97 is developed. The second significant technologies are related to A61Q 19/00 and A61K 8/97. The third significant technologies are related to G01N 33/574 and C12Q 1/68. Table 5 shows the top 3 association rules based on the degree of confidence.

As shown in Table 5, all the degrees of confidence of the 3 association rules have a value of 1. If a technology of C12N 15/82 is developed, a technology of A01H 5/00 is necessarily developed, and if a technology of A61Q 7/00 is developed, a technology of A61K 8/97 is necessarily developed. However, although the degree of confidence between two technologies is 1, the degree of support is very low. Consequently, these technologies are likely to be vacant technologies, which are technologies that are not properly researched or developed.

As shown in Table 6, the first two association rules have a very large value at 51.2 in the degree of lift. If a technology of C12N 15/82 is developed in RFT, a technology of A01H 5/00 is necessarily developed, and if a technology of A01H 5/00 is developed, the probability that a technology of C12N 15/82 being developed is about 70%. In the second association rule, if technologies of C12Q 1/68 and G01N 33/68 are developed, the probability that a technology of G01N 33/574 being developed is about 75%. In the third association rule, if technologies of C12Q 1/68 and G01N 33/574 are developed, the probability that a technology of G01N 33/68 is developed is about 60%. However, the degrees of support are very low, which means that these technologies are not actively developed in RFT.

Table 7 shows the top association rules identified using 491 IPC codes included in 512 RFT-related patent documents. According to Table 7, if the technologies of "separation, production and purification related to plant cells" are developed, the technologies of "proliferation of flowering plants using tissue culture technology" are developed, and also other association rules are identified for RFT.

Table 5. Top association rules using the degree of confidence

X→Y	rank	confidence	support	lift
C12N 15/82→A01H 5/00	1	1	0.0137	51.2000
A61Q 7/00 →A61K 8/97	2	1	0.0176	7.8769
C12N 15/11→C12Q 1/68	3	1	0.0156	15.5152

Table 6. Top association rules using the degree of lift

X→Y	rank	support	confidence	lift
C12N 15/82→A01H 5/00 A01H 5/00 →C12N 15/82	1	0.0137	1.0000 0.7000	51.2000
{C12Q 1/68,G01N 33/68}→G01N 33/574	2	0.0176	0.7500	22.5882
{C12Q 1/68,G01N 33/574}→G01N 33/68	3	0.0176	0.6000	19.2000

Table 7. Top association rules of RFT and detailed technologies thereof

Association rule	IPC codes and technologies	
C12N 15/82→A01H 5/00 A01H 5/00 →C12N 15/82	C12N 15/82	Separation, production and purification related to plant cells
	A01H 5/00	Proliferation of flowering plants using tissue culture technology
{C12Q 1/68,G01N 33/68}→G01N 33/574 {C12Q 1/68,G01N 33/574}→G01N 33/68	C12Q 1/68	Methods of sampling, or inoculating or spreading a sample involving nucleic acids
	G01N 33/68	Immunoelectrophoresis involving proteins, peptides or amino acids
	G01N 33/574	Immunoelectrophoresis for cancer

4. Conclusion

This study was conducted to analyze the patents for radiation-related fusion technologies that are recently becoming more significant, using statistical methods and data mining techniques. RFT, which is a technology of fusing radiation with information technology, biotechnology, nanotechnology and environmental technology, creates new added value by developing radioactive new drugs or radiation healthcare technologies and eco-friendly radiation fusion technologies. In particular, the development of new and high value-added bio-products using the fusion of radiation technology with biotechnology and the development of medical and industrial technologies using bio-material production technology and radiation genomics are

emerging as a key technology of the 21st century bio-tech industry. Thus, as RFT becomes more and more important, it is very meaningful to forecast new technologies by systematically analyzing the relevant patent information.

The findings of this study are summarized as follows:

First, most of the top 10 patents are related to cosmetic development, drug development and food processing. In other words, the technologies used the most in RFT are technologies for "emersion of cosmetics or quasi-cosmetics derived from herbs (e.g. herbal extracts) (A61K 8/97)" and "antineoplastic agents (A61P 35/00)."

Second, in terms of the degree of support, the degree is high when a technology of A61Q 19/08 is first developed and then a technology of A61K 8/97 is developed. The second technology is related to A61Q 19/00 and A61K 8/97 and the third technology is related to G01N 33/574 and C12Q 1/68.

Third, in the top association rules identified using 491 IPC codes included in 512 RFT-related patent documents, if the technologies of "separation, production and purification related to plant cells" are developed, the technologies of "proliferation of flowering plants using tissue culture technology" are developed.

The findings of this study have some limitations because this study targeted only the patents registered in the korean intellectual property office. However, the findings of this study are significant as they have derived basic data for technological forecasting by identifying specific information about core technology factors included in each patent for radiation anti-oxidation technology and also the relationships among technologies, central technologies and patents thereof. Accordingly, it is thought that the findings of this study can be used as basic data for technological forecasting.

5. References

1. Choi J, Kim H, Im N. Keyword network analysis for technology forecasting. Journal of Intelligence and Information System. 2011; 17(4):227–40.
2. Bengisu M, Nekhili R. Forecasting emerging technologies with the aid of science and technology databases. Technological Forecasting and Social Change. 2006; 73(7):835–844.
3. Weaver WT. The delphi forecasting method. The Phi Delta Kappan. 1971; 52(5):267–71.
4. Agami NME, Omran AMA, Saleh M M, Shishiny HEEE. An enhanced approach for trend impact analysis. Technology Forecasting and Social Change. 2008; 75(9):1439–50.
5. Lee SJ, Lee SH, Seol HJ, Park YT. Using patent information for designing new product and technology: Keyword based technology roadmapping. R&D Management. 2008; 38(2):169–88.
6. Jun S, Lee SJ. Emerging technology forecasting using new patent information analysis. International Journal of Software Engineering and its Applications. 2012; 6(3):107–16.
7. Choi J, Hwang YS. Patent keyword network analysis for improving technology development efficiency. Technological Forecasting and Social Change. 2014; 83:170–82.
8. Lee S, Yoon B, Park Y. An approach to discovering new technology opportunities: Keywords-based patent map approach. Technovation. 2009; 29(6–7):481–97.
9. Kho JC. A study on research trend in management of technology using keywords network analysis, Sungkyunkwan University Ph. D. Thesis: Seoul; 2013. p. 752–53.
10. Huh MH. Introduction to social network analysis using R, Freedom Academy Press: Seoul; 2012.
11. Scott JG . Social network analysis(3rd ed.), SAGE: Los Angles; 2012.
12. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987; 20:53–65.
13. Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. New Jersey: John Wiley and Sons; 1990.
14. Everitt BS, Landau S, Leese M, Stahl D. Cluster analysis. 4th ed. New Jersey: Wiley. 2001;13(2):336–42.