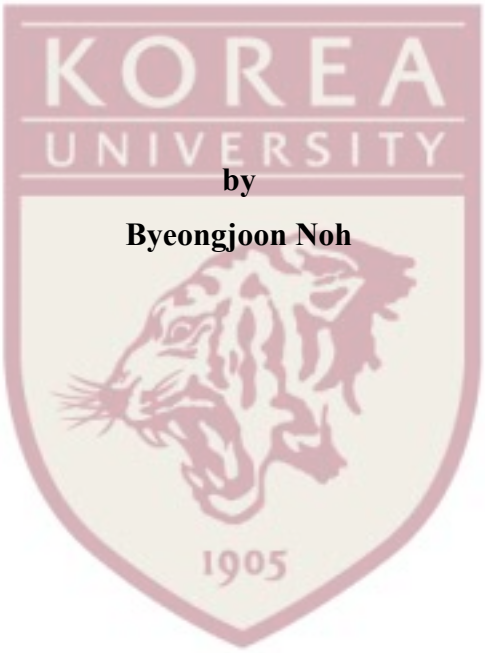
Thesis for the Degree of Master

Study on Keyword-Cube Model with Network Analysis and Association Rule Mining



by

Byeongjoon Noh

Department of Computer and Information Science

Graduate School

Korea University

December, 2016

朴 大 熙 敎 授 指 導

碩 士 學 位 論 文

Study on Keyword-Cube Model
with Network Analysis and Association Rule Mining

이 論文을 理學碩士學位 論文으로 提出함.

2016年 12月 6日

高 麗 大 學 校 大 學 院

컴퓨터 情報學科

盧 炳 峻

盧炳峻의 理學 碩士學位 論文 審査를 完了함.

2016年 12月 6日

委員長 방 대 희 (印)

委 員 전 태 규 (印)

委 員 정 동 화 (印)

ABSTRACT

Study on Keyword-Cube Model with Network Analysis and Association Rule Mining

Byeongjoon Noh
Advised by Prof. Daihee Park
Major in DB & Data mining
Dept. of Computer and Information Science
Graduate School
Korea University

Recently, the foot-and-mouth disease (FMD) can spread rapidly, causing the social issues such as high mortality rates and considerable economic damage. Therefore, it is necessary to analyze these issues incurred from such livestock diseases such as repercussion effects and response actions.

In this thesis, we propose a keyword analytical system using online news articles for scrutinizing a variety of social issues caused from FMD outbreaks. Based on keywords extracted from online news, the proposed system performs the multidimensional analysis by constructing keyword-cube model to discover the hidden knowledge incurred from FMD outbreaks with

diverse levels of abstraction. Moreover, the relationships among issues are analyzed by applying co-occurrence keywords network and association rule mining to obtain the comprehensive information of them. We validate the feasibility and applicability of the proposed analytical system by implementing and applying it to FMD outbreaks from November 2010 to December 2011 in South Korea.

.

Keywords: *Online news analysis, Keyword analysis, Data cube, OLAP operation, Association rule mining, Network analysis, Livestock disease analysis*

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

Recently, the Foot-and-Mouth (FMD) can spread rapidly, resulting in high mortality rates and adverse social issues such as agricultural-livestock industries, water pollution and funding problem [1, 2]. Therefore, it is necessary to analyze these issues incurred from such livestock diseases in the various perspectives. To date, much research has been conducted to prevent livestock disease from spread [3-8], and to analyze the economic effects from livestock disease [9-11] in the various fields.

However, most of the current livestock disease analysis system [3-11] mainly focus on prevention from spread or confined analysis for only economic effects. Because highly inspective livestock disease, such as FMD, causes a variety of social issues, the analysis must be conducted in various perspectives such as economic, environment and policy.

To address these challenges, in this thesis, online news articles are used to scrutinize the various issues from livestock disease. With propagating the information in fast, the online news articles not only contain the variety social issues, such as society, culture and economy, but also have reliability and well-organized structures [12, 13]. In addition, the text data processing techniques can facilitate discovery of knowledge and visualization such as

topic-modeling [14], network analysis, natural language processing and sentimental analysis [15, 16].

In this thesis, we propose a keyword analytical system using online news articles for scrutinizing a variety of social issues cased from FMD outbreaks. Based on keywords extracted from online news, the proposed system performs the multidimensional analysis by constructing keyword-cube model to discover the hidden knowledge incurred from FMD outbreaks with diverse levels of abstraction. Moreover, the relationships among issues are analyzed by applying co-occurrence keywords network and association rule mining to obtain the comprehensive information of them. The issues, which are analyzed in this study, are largely grouped into two response actions and three repercussion effects. The response actions are; 1) direct actions which taken at the site such as stamping out and burying; 2) indirect actions which taken at the command center such as vehicle control. The repercussion effects are; 1) the economic effects resulting in the high mortality rates which have badly affected in the various fields as well as livestock industries; 2) the political effects such as coping with crisis and funding problem; 3) the environment effects resulting in soil or groundwater contamination from buried cadaver.

To the best of our knowledge, this is the first report of scrutiny for the

various issues from livestock disease by using online news articles with OLAP operations, network analysis and association rule mining. We confirm the feasibility and applicability of the proposed keyword analytical model by implementing and applying it to the analysis for repercussion effects and responding actions of FMD outbreaks in South Korea.

The remainder of this paper is organized as follows. Section 2 introduces research backgrounds and related work on livestock disease analysis, OLAP system, network analysis, association rule mining method. Section 3 presents the proposed keyword-analytical system using OLAP operation, network analysis and association rule mining, in detail. The experimental results are demonstrated in Section 4. Finally, concluding remarks are given in Section 5.

2. RELATED WORK

2.1. Livestock Disease Analysis

The high infectious livestock diseases have increased the threat to human life, raising the considerable economic damage. Therefore, it is necessary to analyze the livestock diseases to minimize the damages from it. Much research has been conducted in terms of pathology [3, 4], prevention from spread [5-8] and economic effects [9-11]. In prevention perspective, the authors in [7] analyze extensive long-term the highly pathogenic avian influenza (HPAI) data by using OLAP operations and sequential pattern mining. The authors in [8] construct direct/indirect HPAI spread network based on the relationships between farms in South Korea, and provide an optimal prevention solutions by analyzing network centrality. In economic effects analysis perspective, the authors in [10] analyze the economic impact of FMD in terms of evaluation for milk yield in Pakistan.

Unlike other research approaches, we analyze the various responding actions and repercussion effects as well as economic effects from livestock disease by using the online news articles.

2.2. On-Line Analytical Processing System

Data warehouses and OLAP systems are based on a multidimensional data model by representing data in form of a data cube. OLAP is useful for an integrated decision-making approach, such as business intelligence, financial reporting and knowledge discovery, for the multi-dimensional data analysis [17, 18]. To date, an extensive variety of studies has been conducted in the various fields by using OLAP system. Especially, in text data processing perspective, the authors in [19] describe a framework which combined OLAP and information networks to deal with bibliographic data. They conduct the comparison between traditional OLAP and OLAP on information networks. The authors in [20-22] construct the data cube based on text-heavy documents to analyze and summarize unstructured data.

In this thesis, we design a data cube model, called *keyword-cube*, using keywords which extracted from online news articles. Based on keyword-cube model, multidimensional analysis is performed by using OLAP operations to discover the repercussion effects of FMD, and to assess the responding actions with varying levels of abstraction.

2.3. Network Analysis

Network analysis is useful to explore structural properties of a set of items and the connections between them, called nodes and edges, respectively. With a nice graphical representation, network analysis have widely used in various fields such as medicine [23], livestock disease analysis [8, 24], bibliography [25, 26], social science [27].

In text data processing perspective, much research conducted to discover and obtain the structure of documents. For example, the authors in [25] examine the research patterns and trends of *RecSys* (Recommendation System) in China. In this study, co-occurrence word network is applied to analyzing correlation among the extracted keywords by using degree of centrality and density. The authors in [26] analyze headlines' keywords and words relationship in online Chinese news by using word network with the various text mining techniques. To construct the co-occurrence word network (word network), words and adjacent relations are considered as nodes and edges, respectively.

In this thesis, we construct the co-occurrence keywords network to examine the relationships between keywords with varying levels of abstraction.

2.4. Association Rule Mining

Association rule mining (ARM) aims in uncovering interesting and useful patterns in a large transactional database [17, 28]. Association rule could be represented in the form of "$X \rightarrow Y$", where X and Y are able to be defined as itemsets.

Association rules are useful in analyzing and predicting customer behavior and it is also important for decision-making tasks. Much research has been conducted in the various fields such as business intelligence [29], industrial maintenance [30] and medicine [31, 32]. Especially, in text data processing perspective, the authors in [33] propose the association rule based hierarchical clustering method (ARBHC) to obtain information in a more efficient manner by using *apriori* algorithm, as one of the commonly used association rule algorithm, and hierarchical clustering algorithm. The authors in [34] propose topic discovery and trend prediction framework to analyze research patterns in fields of data mining and machine learning by using clustering, topic-modeling and association rule mining.

In this thesis, *apriori* algorithm is employed to discover association rules and to analyze the relationship between repercussion effects and responding actions with diverse levels of abstraction.

3. MATERIALS AND METHODS

3.1. The Overall Architecture of Proposed Method

The **Figure 1** illustrates an overall concept of keyword analytical system which is proposed in this thesis. As you can see in **Figure 1**, the proposed system consists of five layers; 1) data source layer; 2) preprocessing layer; 3) data storage layer; 4) keyword analytics layer; 5) visualization layer. In first layer, the online news articles are collected from web portal site. The collected news articles are preprocessed and keywords are extracted by using topic-modeling algorithm. Thus, keyword taxonomy is constructed to analyze keywords in a variety of perspectives. In third layer, we use the data warehouse for the effective storage of large-scale articles. In the keyword analytics layer, we conduct the analysis by applying OLAP operations, network analysis and association rule mining. Finally, we represent the results of analysis on dashboard which supports decision-maker.

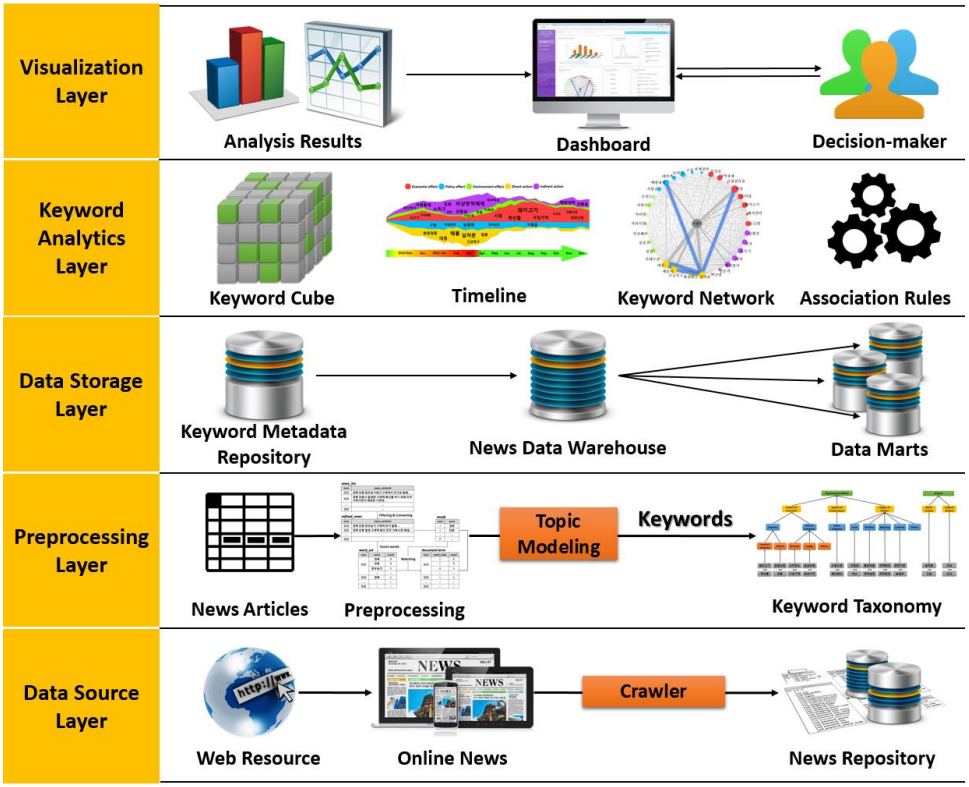


Figure 1. The overall architecture of proposed method

3.2. Data Sources

In data source layer, as the first layer of the proposed system, online news articles are collected from web portal site. We use the '*beatifulsoup*' library in *Python*, as collection tool. The collected raw data contains several attributes such as title, time, author and contents from November 2010 to December 2011. Thus the articles, which include the term of "foot-and-mouth disease" in '*contents'*, are selected as dealing with issues of FMD. As a result, **Figure 2** presents the distributions for the number of FMD outbreaks and FMD news in collection period [35].

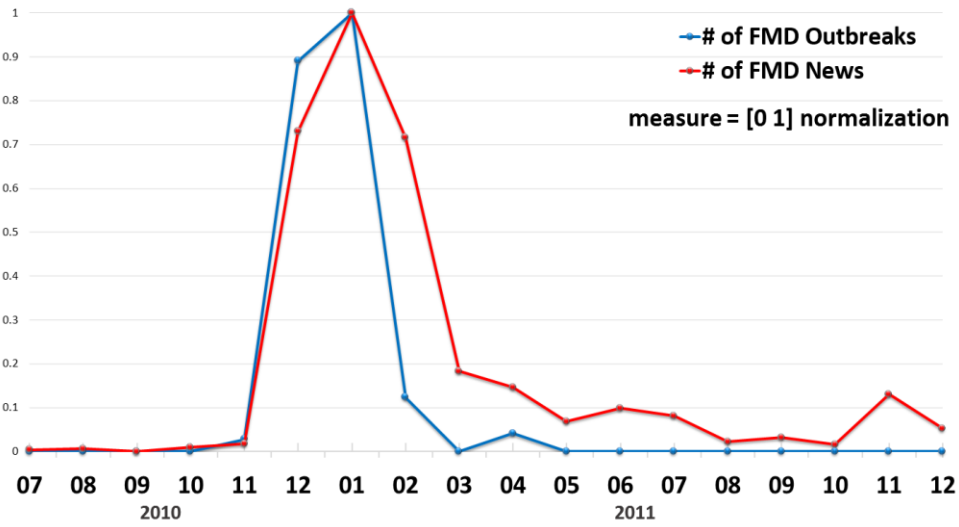


Figure 2. Result of the collected news articles

3.3. Pre-processing for Analysis

In this section, we describe the preprocessing steps in detail. Since the news articles are a form of unstructured information which is text-heavy, it is necessary for the collected articles to be preprocessed. As the first step, stop-words which have no information are removed. Next, the words are filtered and converted to extract keywords more exactly. The rules of filtering and converting are followed in **Table 1**. After applying the rules, the nouns are exacted by using '*KoNLP*' package, as one of the natural language processing toolkits for Korean, in R program [36].

Table 1. The rules of filtering and converting

No	Rule of Filtering & Converting	Objects	Result
1	Remove the names of author and reporter included in each news article	Reporter : ** K** news Copyright all rights reserved	[Removed]
2	Remove the advertisements phrase	Take an extra 30% off	[Removed]
3	Convert the error nouns extracted from *KoNLP*	occurs occurred occurring	occurrence
4	Unify the name of location	Gyeongsangbuk-do Gyeongbuk province North Gyeongsang	Gyeongbuk

In second step, the Latent Dirichlet Allocation (LDA), as one of the most commonly used topic-modeling algorithm, is used to extract the keywords from filtering and converting articles. As the input data for LDA, document sets (articles) are represented in form of document-term matrix [37]. As you can be seen in **Table 2**, the rows and columns in this matrix correspond to the documents and terms, respectively. The entry m_{ij} indicates how often the j^{th} term occurred in the i^{th} document. Based on this matrix, we extract 20 keywords for each month (from Nov.2010 to Dec. 2011).

As final step, we design the keyword taxonomy which categorized into word sense. The purpose of constructing keyword taxonomy is to use it as the hierarchy for the keyword dimension in a keyword-cube to analyze keywords with the various levels of abstraction. In keyword taxonomy, the keywords are largely grouped into two parts such as repercussion effects and responding actions. The repercussion effects consist of economic, environment and policy effects, and the responding actions consist of direct actions and indirect actions as **Figure 3**.

Table 2. Example of *document-term* matrix

Terms	Documents				
	Doc_1	Doc_2	Doc_3	…	Doc_n
apple	3	0	4		8
livestock	1	5	7		7
disease	2	2	13	…	8
…	…	…	…		…
emergency	0	4	6		1

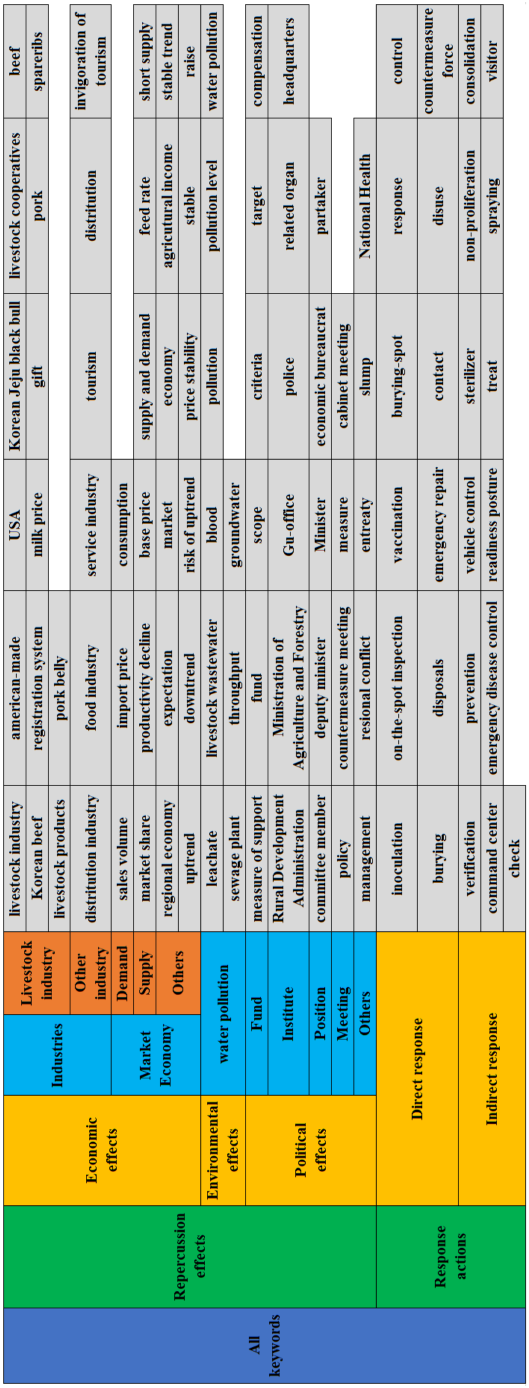


Figure 3. The extracted keywords and keyword taxonomy

3.4. Keyword-Cube Model

3.4.1. Star Schema

A multidimensional model is one of the commonly used data model for a data warehouse and OLAP tools with representing a form of data cube. A data cube allows data to be modeled and views in multiple dimensions. It is defined by facts and dimensions. The fact table contains the names of the facts, or measures, as well as keys to each of the related dimension tables. Dimensions are the perspectives or entities with respect to which a user wants to keep records.

In this thesis, a star schema is used for the multidimensional analysis model. We defined four dimensions; keyword, time, disease and livestock. In fact table, we apply "*frequency*", frequency of keyword, as the measurement. **Figure 4** demonstrates the star schema for the proposed model.

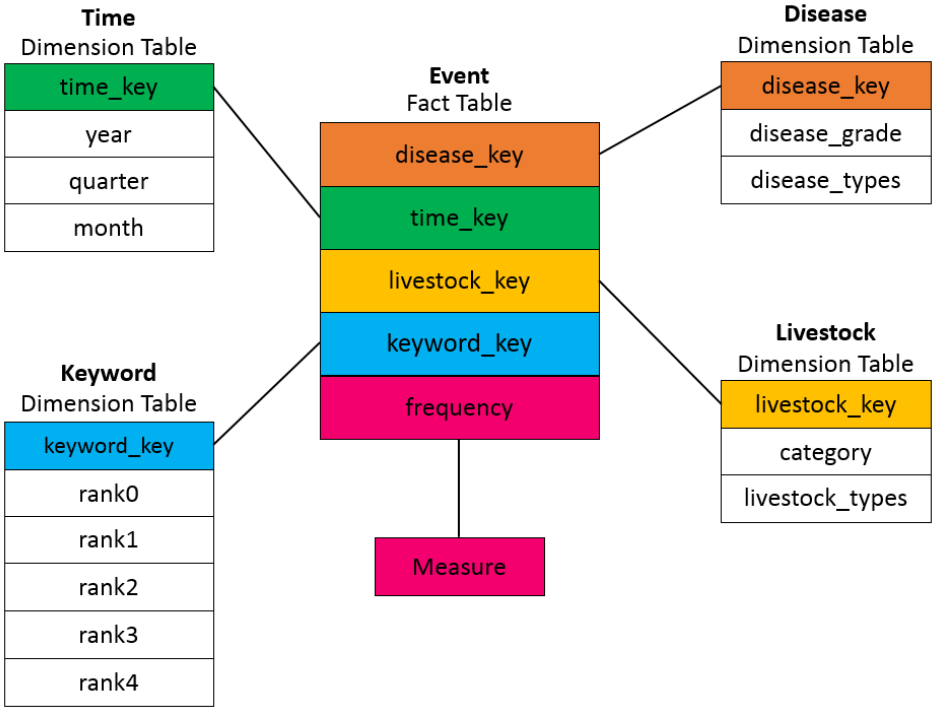


Figure 4. Star schema for keyword cube model

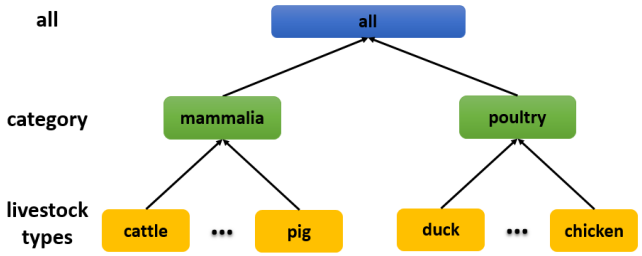
3.4.2. Concept Hierarchy and OLAP Operations

A concept hierarchy defines a sequence of mappings from a set of low-level concepts to high-level and more generalized concepts. It allows data to be managed at varying levels of abstraction. As you can see in **Figure 5**, we design the concept hierarchies for constructing *keyword-cube* model. The attributes of time dimension are organized as **Figure 5 (a)** such as month, quarter and year. The concept hierarchy for keyword dimension, as the core of *keyword-cube*, is based on hierarchical tree, from the lowest level (*keyword*, *rank4*) to the highest level (*rank0*) as **Figure 5 (d)**. The main idea of *keyword-cube* is to use the keyword taxonomy as the hierarchy for the keyword dimension to analyze keywords with a variety levels of abstraction. Similarly the concept hierarchies for livestock and disease are presented in **Figure 5 (b)** and **(c)**, respectively.

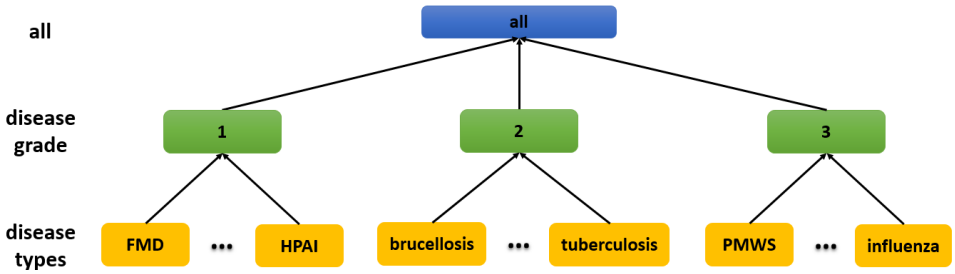
By using OLAP operations which provide users with the flexibility to view data from different perspectives, the useful information is retrieved from the news data warehouse. Furthermore, OLAP operations have characteristics of a user-friendly environment for interactive data analysis. In other words, it is possible to construct a flexible and extensible analysis system. To scrutinize keywords in multiple perspectives from *keyword-cube*, we use the various

OLAP operations such as *roll-up*, *drill-down*, *slice* and *dice*. The *roll-up* operation performs aggregation on a data cube, either by climbing a concept hierarch for a dimension or by dimension reduction. The *drill-down* operation is the reverse of the *roll-up* operation; it navigates from less detailed data to more detailed data. The *slice* operation selects one dimension of a given cube resulting in a sub-cube. The *dice* operation defines a sub-cube by performing a selection on two or more dimensions [17].

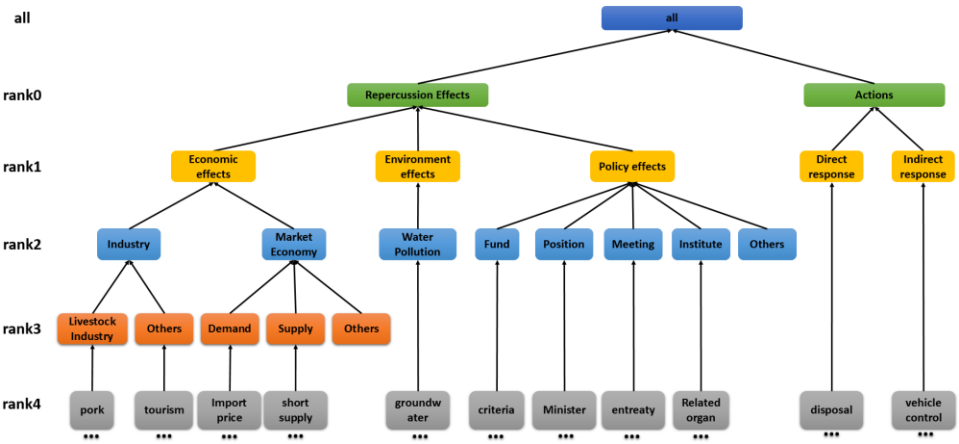
(a) Concept hierarchy of time dimension



(b) Concept hierarchy of livestock dimension



(c) Concept hierarchy of disease dimension



(d) Concept hierarchy of keyword dimension

Figure 5. Concept hierarchy of (a) time dimension; (b) livestock dimension; (c) disease dimension; (d) keyword dimension

3.5. Network Analysis

The structure of network is one of the widely used to explore and discover the structural properties in which certain words appear at the same time and grasping the structural relation of documents and words in field of bibliography. In general, network is represented in form of matrix, and there exist two types for representing network: two-mode matrix and one-mode matrix [39]. Two-mode matrix has the rows and columns index from the different set of entities. In contrast, one-mode matrix has the rows and columns index from the same set of entities. It enables a user to understand the whole relationship easier among the objects than two-mode network as well as intuitive graphical display [40].

In order to construct the co-occurrence keywords network, as the proposed analytical methodology, we have to transform document-keyword matrix (network) into keyword-keyword matrix as Table 3. For example, as **Table 3 (a)**, $keyword_1$ and $keyword_2$ are appeared in two documents (Doc_3 and Doc_4), simultaneously. Then, the value of $keyword_1$ - $keyword_2$ is 2 as **Table 3 (b)**. In our network, the size of node and edge means the frequency of keyword and co-frequency of keywords pair, respectively. We used '*igraph*' package in R to construct network [41].

Table 3. The method of converting into adjacency matrix

	$keyword_1$	$keyword_2$	$keyword_3$	$keyword_4$
Doc_1	2	0	1	3
Doc_2	0	1	0	2
Doc_3	1	1	0	0
Doc_4	4	1	17	0
Doc_5	0	1	1	4

(a) Document-keyword matrix

↓

	$keyword_1$	$keyword_2$	$keyword_3$	$keyword_4$
$keyword_1$	.	.	.	.
$keyword_2$	2	.	.	.
$keyword_3$	2	2	.	.
$keyword_4$	1	2	2	.

(b) Keyword-keyword adjacency matrix

3.6. Association Rule Mining

Association rule mining aims at uncovering interesting and useful patterns in large data sets. It consists of first determining frequent itemsets (a set of items, such as A and B, satisfying a minimum support threshold or percentage of the task-relevant tuples) from which strong association rules in the form of $A \rightarrow B$ are generated. These rules also satisfy a minimum confidence threshold (a prespecified probability of satisfying B under the condition that A is satisfied). Associations can be further analyzed to uncover correlation rules, which convey statistical correlations between A and B.

In our experiment, *apriori* algorithm, as one of the commonly used ARM algorithm, is used to discover association rules among keywords of FMD such as repercussion effects and responding actions with multiple levels of abstraction. With usage of '*arules*' package in R program, the extracted keywords set for each month are used as transaction as **Table 4**.

Table 4. Data set for association rule mining based on keywords

Date	TID	Transaction
Nov. 2010	Tr_1	{pork, command center, agriculture income, livestock cooperatives, response, countermeasure force, emergency disease control, livestock product, beef, measure, entreaty, Rural Development Administration, treat, check, police, partaker}
Dec. 2010	Tr_2	{Gu-office, pork, Ministration of Agriculture and Forestry, prevention, command center, vehicle control, market, disposals, visitor, livestock products, market share, emergency disease control, consolidation, burying}
Jan. 2011	Tr_3	{vaccination, non-proliferation, spareribs, regional economy, disposals, countermeasure meeting, beef, burying, prevention, command center, livestock products, sewage plant, visitor, Ministration of Agriculture and Forestry, sterilizer, emergency repair, consolidation, emergency disease control}
…	…	…
Dec. 2011	Tr_{14}	{livestock products, sales volume, livestock industry, consumption, supply and demand, sewage plant, non-proliferation, import price, pork, measure of support, water pollution, command center, readiness posture, contact, tourism, regional economy, stable trend, distribution}

4. EXPERIMENTS AND RESULTS

In this section, we describe the experimental design and perform the analysis to test the feasibility and applicability of the proposed system.

4.1. Experimental Design

In this section, we describe the experimental design for proposed online news analytical system using OLAP, network analysis, association rule mining. First, in analysis utilizing OLAP, we conduct the analysis of changes for diverse repercussion effects and response actions according the time. The analysis for repercussion effects is expected to enable deriving not only the economic adverse effects of the outbreaks of FMD appearing in the livestock industry but also the other issues such as chained repercussion effects on other industrial fields. In addition, analysis for response actions is expected to enable the assessment of whether or not appropriate response actions were taken according to the severity of FMD. Finally, the frequencies of keywords according to time will be expressed in the form of *Themeriver* to examine the analysis of the trends of major social issues related to the outbreaks of FMD [43].

In the analysis using keywords network and association rule mining, we define the time of outbreaks of FMD as three periods such as 'early period', 'serious period' and 'termination period' (see **Table 5**). In the keywords network analysis, we perform the analysis for keywords pairs of top-5 co-frequencies with levels of abstraction by constructing co-occurrence keywords network for each period.

Table 5. Defined of periods for FMD outbreaks

	Early period	**Serious period**	**Termination period**
Term	Nov.2010 ~Dec.2010	Jan.2011 ~ Mar.2011	Apr. 2011 ~ Dec. 2011
# of FMD outbreaks	415	1148	83

In association rule mining, the case studies are conducted only for those rules that appear in the form of {*action*} → {*effect*} among diverse rules. **Table 6** shows that input parameters and output for ARM in each period

Table 6. Input parameters and outputs of ARM in each period

	parameter	In early period	In serious period	In termination period
Input	Transaction number	$Tr_1,\ Tr_2$	$Tr_3 \sim Tr_5$	$Tr_6 \sim Tr_{14}$
	Minimum Support	0.5	0.3	0.1
	Minimum Confidence	1.0	1.0	1.0
Output	# of the all rules	28	128	243

4.2. Analysis Results using OLAP

In this section, we describe the multidimensional analysis results which can be obtained by choosing a dimension within the viewpoint that a decision-maker is attempting to scrutinize, and also by adjusting the levels of abstraction. Prior to analysis, we can observe that the various issues were arisen from FMD outbreaks according to time slots, by using following formula (1), which is a query statement with a symbolic meaning:

* *Dice* for (disease=*"FMD"*) and (livestock =*"all"*)
and (keyword=*"rank0"*) and (time=*"*each *year"*)
and (measure=*"frequency"*) (1)

From formula (1), **Figure 6** presents the intensive repercussion effects and response actions during the whole periods (Nov. 10 ~ Dec. 11). In order to understand the results of the detailed analysis for scenario I and II, we conduct *drill-down* and *dice* operations for the time and keyword dimensions.

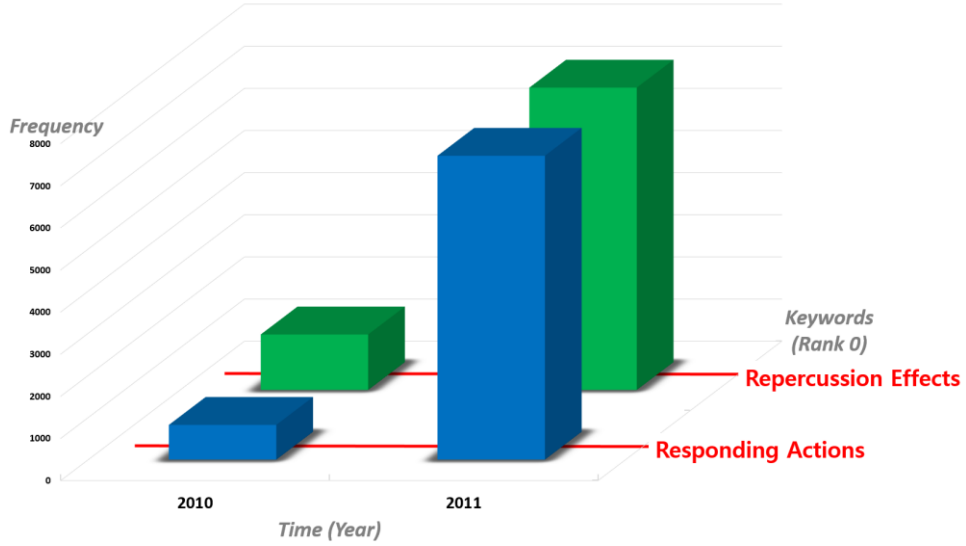


Figure 6. Result of keyword frequency from 2010 to 2011

4.2.1. Scenario 1 – Analysis for Time-Repercussion Effects

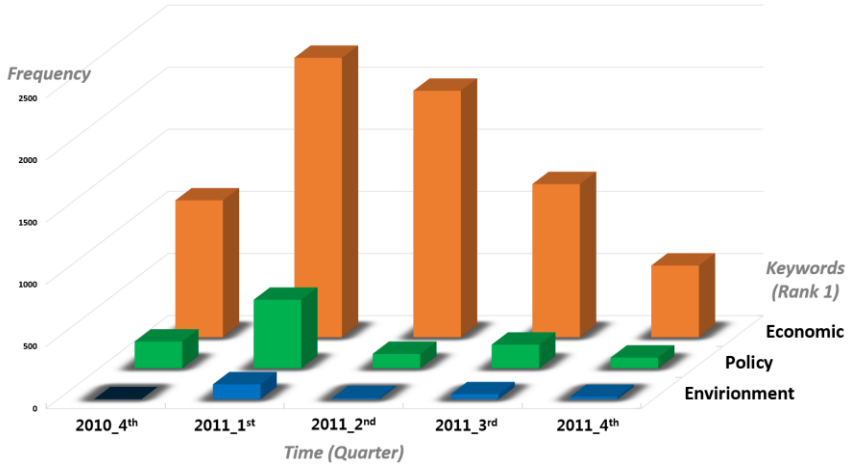
We conduct formula (2) from **Figure 6** to analyze keywords in terms of repercussion effects.

* *Drill down* on time (*year* to *quarter*)
* *Drill down* on keyword (*rank1* in *repercussion*)
* *Dice* for (disease="*FMD*") and (livestock ="*all*")
and (keyword="each *rank1* in *repercussion*") (2)
and (time="each *quarter*") and (measure="*frequency*")

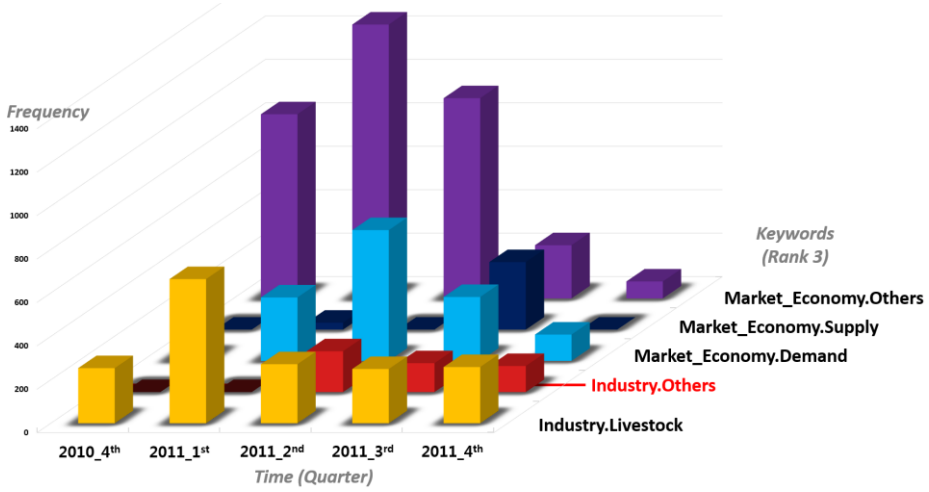
Figure 7 (a) indicates the graph for frequency of repercussion effects in each quarter. We can observe that frequency of economic effects is higher than others in the whole periods. In order to perform the detailed analysis for economic effects, we drilled down from *rank1* level to *rank3* level for the keyword dimension, as illustrated in formula (3).

* *Drill down* on keyword (*rank3* in *economic*)
* *Dice* for (disease="*FMD*") and (livestock ="*all*")
and (keyword="each *rank3* in *economic*")
and (time="each *quarter*") and (measure="*frequency*") (3)

Figure 7 (b) shows that there exist adverse influence in fields of livestock industry during FMD outbreaks. On the other hand, the keywords for other industries appear in only after termination of FMD outbreaks.



(a) Result of repercussion effects for *each quarter* in *rank1* level



(b) Result of repercussion effects for *each quarter* in *rank3* level

Figure 7. Result of repercussion effects for each *quarter*

To conduct the analysis for other industries after termination of FMD outbreak, we again drilled down from *quarter* level to *month* level for time dimension and from *rank3* level to *rank4* (*keyword*) level for keyword dimension, as following formula (4).

* *Drill down* on time (*quarter* to *month*)
* *Drill down* on keyword (*rank4* in *industry.others*)
* *Dice* for (disease="*FMD*") and (livestock ="*all*")
and (keyword="each *rank4* in *industry.others*")
and (time="each *month* in *2011_3rd* and *2011_4th* ")
and (measure="*frequency*") (4)

In **Figure 8**, we can observe that keywords for the various industries, such as tourism industry, distribution industry and food industry, appear after termination of FMD. It means that there exist negative influences in these industries. Especially, regional festivals would frequently be cancelled due to FMD outbreaks, causing a series of damage in tourist service industries such as restaurants, accommodation facilities and transportation companies.

In short, there exist the various repercussion effects incurred from FMD outbreaks. Especially, after termination of it, the keywords for economic effects appear in high frequency. In detailed analysis, FMD outbreaks cause the damages in fields of tourism industry as well as livestock industry.

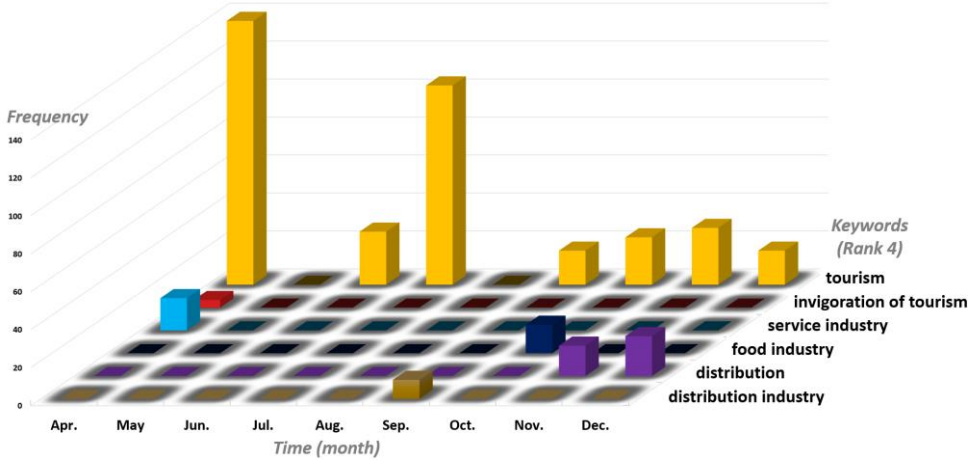


Figure 8. Result of other industries for Apr.11 ~ Dec.11 in *rank4*

4.2.2. Scenario 2 – Analysis for Time-Responding Actions

In order to analyze the response actions according to the time slots, we apply formula (5) from **Figure 6**.

* *Drill down* on time (*year* to *quarter*)
* *Drill down* on keyword (*rank1* in *response*)
* *Dice for* (disease=*"FMD"*) and (livestock =*"all"*)
and (keyword=*"*each *rank1* in *response"*)
and (time=*"*each *quarter"*) and (measure=*"frequency"*) (5)

As you can be seen in **Figure 9**, we can observe that there exist keywords for direct/indirect actions with the high frequencies during FMD outbreaks (red line) and keywords for indirect actions after termination of it (blue line). From former, we will analyze whether proper actions had been taken or not. From letter, we will analyze why indirect actions had been taken in this period.

To perform analysis for first-case (red line) in detail, we drilled down from *quarter* level to *month* level for time dimension and from *rank1* level to *rank4* level for keyword dimension, as indicated in formula (6).

* *Drill down* on time (*quarter* to *month*)
* *Drill down* on keyword (*rank4* in *response*)
* *Dice for* (disease=*"FMD"*) and (livestock =*"all"*)
and (keyword=*"*each *rank4* in *direct_response"*)
and (time=*"*each *month in 2010_4*th and *2011_1*st*"*)
and (measure=*"frequency"*) (6)

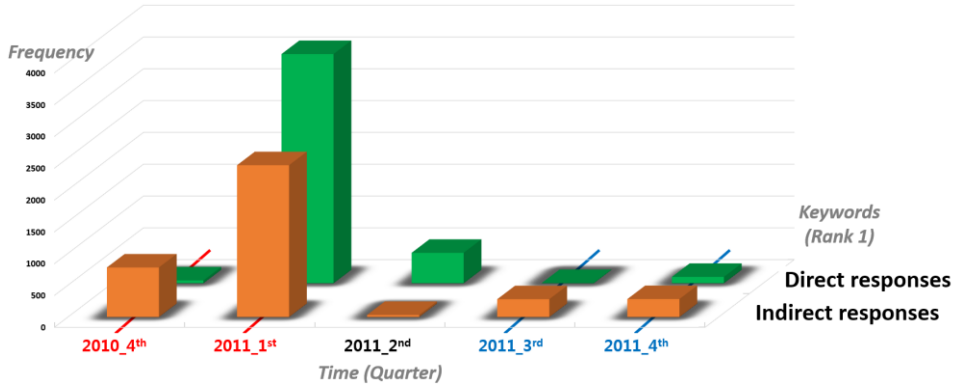
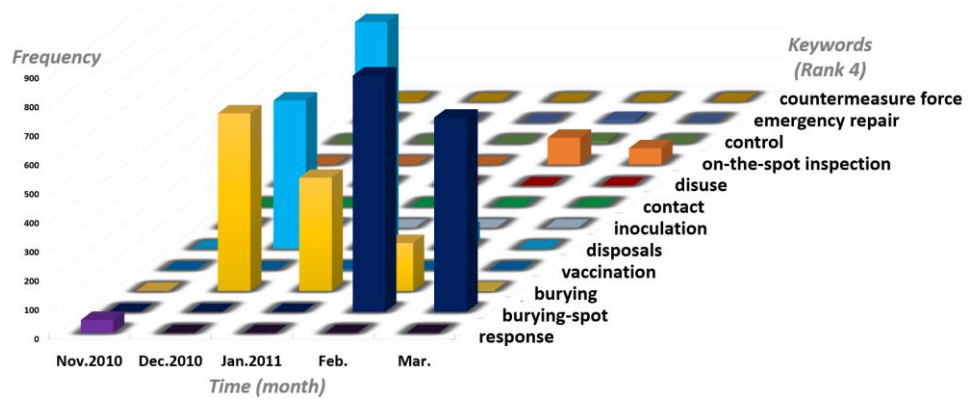


Figure 9. Result of response actions for each *period*

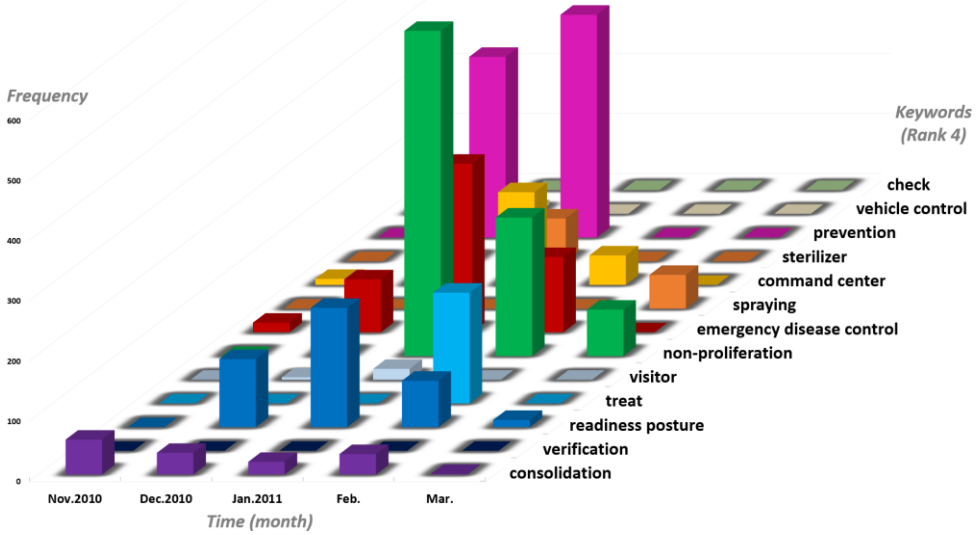
Figure 10 (a) and **(b)** present the graph for direct and indirect actions from Nov. 2010 to Mar. 2011, respectively. In **Figure 10 (a)**, we can observe that the proper responses would be taken well such as burying, vaccination and disposals based on Response Manual for FMD. In **Figure 10 (b)**, we can also identify the keywords which mean the proper indirect actions such as vehicle control, command center and emergency preparedness. Especially, these keywords appear with highest frequencies on January 2011 when the FMD occurred the most serious.

In order to conduct the analysis for second-case (blue line) in detail, we also drilled down from *quarter* level to *month* level for time dimension and from *rank1* level to *rank4* level for keyword dimension, as indicated in formula (7).

* *Drill down* on time (*quarter* to *month*)
* *Drill down* on keyword (*rank4* in *response*)
* *Dice for* (disease=*"FMD"*) and (livestock =*"all"*)
  and (keyword=*"*each *rank4* in *indirect_response"*)
  and (time=*"*each *month in 2011_3rd* and *2011_4th"*)
  and (measure=*"frequency"*) (7)



(a) Result of direct response in *keyword* level during FMD outbreaks



(b) Result of indirect response in *keyword* level during FMD outbreaks

Figure 10. Result of response during FMD in *rank4*

Figure 11 presents the graph for indirect response actions after termination of FMD. In **Figure 11**, we can observe that keywords, which mean command center and preventing, are emerged with high frequencies. Through this result, there are the efforts to protect recurrence of FMD in winter season.

In short, we conducted the analysis for response actions with OLAP operations. The various direct/indirect actions, based on Response Manual for FMD, were taken actively during FMD outbreaks. After termination of FMD, the indirect responses were taken to prevent recurrence of it in winter season.

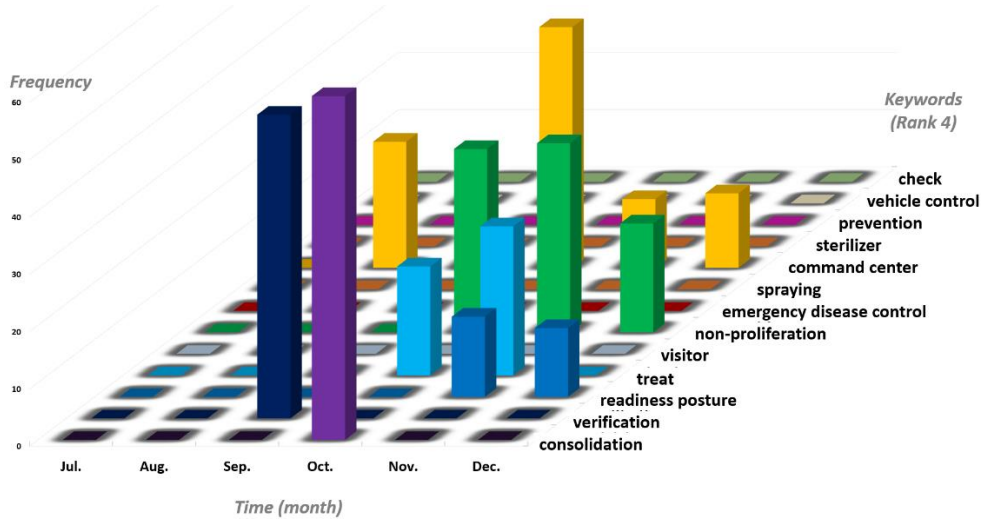


Figure 11. Result of indirect responses after termination of FMD

4.2.3. Keyword Trend Analysis using Timeline

In this section, we perform the trend analysis for changes in social issues such as repercussion effects and response actions from FMD in the entire periods. As you can see in **Figure 12**, a timeline was constructed in the form of *Themeriver*. To sum up, we can observe that the economic effects became issues in the entire periods of the outbreak of FMD and that, in particular, economic effects intensively appeared after termination of it. The keywords for policy effect consistently appeared during the entire periods. Keywords indicating responding parties such as the 'Ministry of Agriculture and Forestry' and 'Gu-office' appeared during FMD outbreaks, whereas keywords indicating ex post facto countermeasures such as 'support fund' appeared after termination of FMD. Meanwhile, the keywords for economic effects would not focus on social issues in same manner as results of OLAP.

In response actions perspectives, keywords related to direct responses such as 'burying' and 'disposals' intensively appeared during FMD outbreaks. As indirect responses, keywords such as 'command center' and 'emergency disease control system' appeared during FMD occurrences. Unlike direct responses, some keywords, which related to the prevention of recurrence of FMD in winter season, appeared as indirect responses.

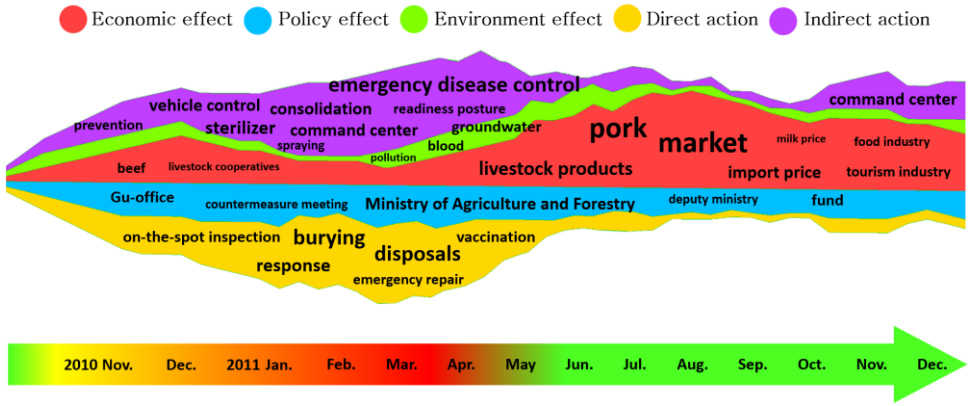


Figure 12. Keyword timeline in each month

4.2.4. Summary and Discussions

In this section, we constructed a keyword-cube model to scrutinize the various social issues such as repercussion effects and response actions according to time by using OLAP operations. This is considered helpful for livestock disease administrators' decision making and devising policy.

However, the proposed keyword-cube model has limitations in identifying the relationships between keywords such as {*effect*}-{*response*}, {*response*}-{*response*} and {*effect*}-{*effect*}. To overcome these limitations, network analysis is performed by constructing co-occurrence keywords network.

4.3. Analysis Results using Keyword Network

4.3.1. Analysis in Early Period

As you can be seen in **Figure 13**, keywords network in early period shows that keywords indicating direct/indirect responses based on *response manual for FMD* were emerged such as ‘disposals’ and ‘emergency disease control’. In repercussion effects perspectives, some keywords, which related to economic and policy effects, were identified such as ‘pork’, ‘livestock products’, ‘police’ and ‘Rural Development Administration’. However, keywords, which related to environmental effects, have not appeared as social issues yet.

In co-frequency perspective, we can observe that co-frequencies for {*direct response*}-{*economic effect*} and {*direct response*}-{*policy effect*} is high as shown in **Table 7**. In detail, we can obtain the fact that the extreme measures such as the disposals had serious adverse effects on the economy of livestock industry due to policy for the prevention of the spread of FMD and further outbreaks.

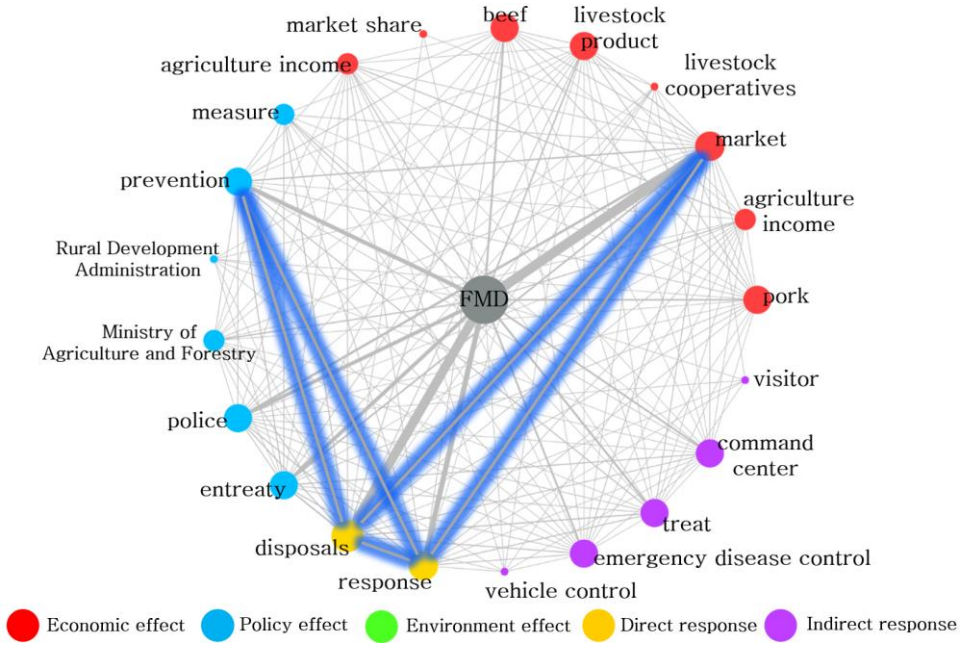


Figure 13. Keywords network in early period

Table 7. Top 5 co-frequencies in early period

No.	Co-keyword	Co-frequency
1	market – disposals	180
2	market – response	108
3	measure – response	107
4	measure – disposals	89
5	disposals – response	76

*** Do not consider the co-frequency of 'FMD'-[keyword]**

4.3.2. Analysis in Serious Period

Figure 14 shows the keywords network in serious period. In Figure 14, we can observe that keywords indicating direct/indirect response were emerged intensively such as ‘disposals’, ‘burying’ and ‘non-proliferation’. Through these keywords, it can be known that FMD were seriously infected, and the various direct response were carried out to prevent further outbreaks and spread. In addition, as the FMD persisted for a long time, a variety of repercussion effects such as economic, policy and environment appeared such as ‘pork ribs’, ‘countermeasures meeting’ and ‘leachate’ (see **Table 8**).

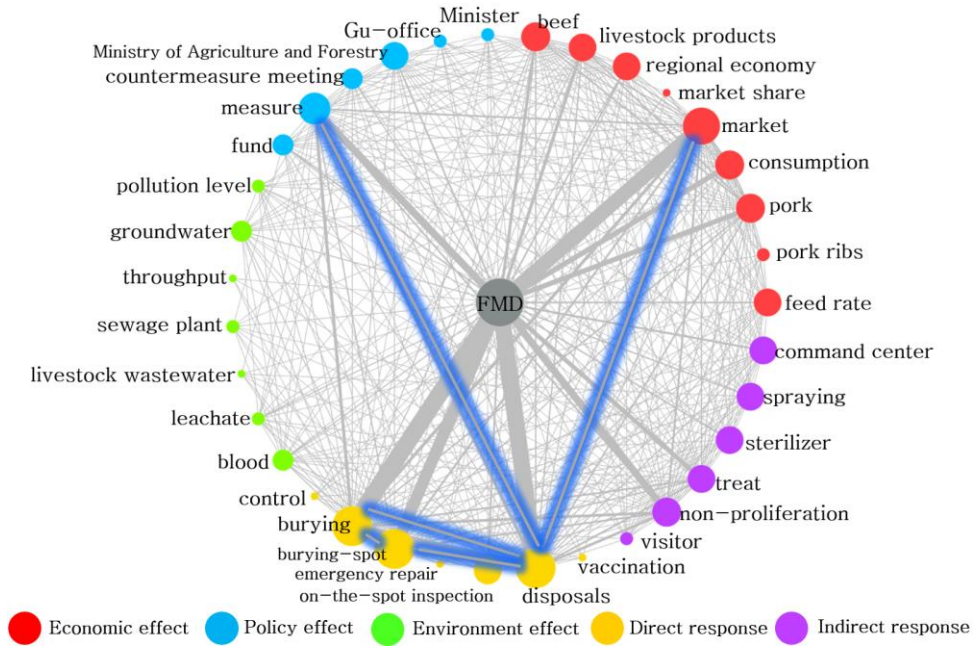


Figure 14. Keywords network in serious period

Table 8. Top 5 co- frequencies in serious period

No.	Co-keyword	Co-frequency
1	burying – disposals	708
2	disposals – burying-spot	503
3	burying-spot – disposals	337
4	market – disposals	268
5	measure – disposals	255

*** Do not consider the co-frequency of 'FMD'-[keyword]**

4.3.3. Analysis in Termination Period

The keywords network in termination period is shown in **Figure 15**. **Figure 15** shows that keywords indicating responses disappeared and diverse repercussion effects were emerged. In particular, many keywords related to economic repercussion effects appeared. As a result, it is confirmed that the price of pork was lowered due to the FMD outbreaks and livestock market was adversely affected. In addition, other industries such as tourism, distribution and service industries are also affected as well as livestock industry. As shown in **Table 9**, the keywords related to the economic repercussion effects is high in termination period. Especially, it can be seen that there was a negative effects on the market economy in field of the livestock industry, such as the decrease in the production of pork and consumption in the market. Meanwhile, it is confirmed that there were some indirect responses to prevent FMD outbreaks in winter season.

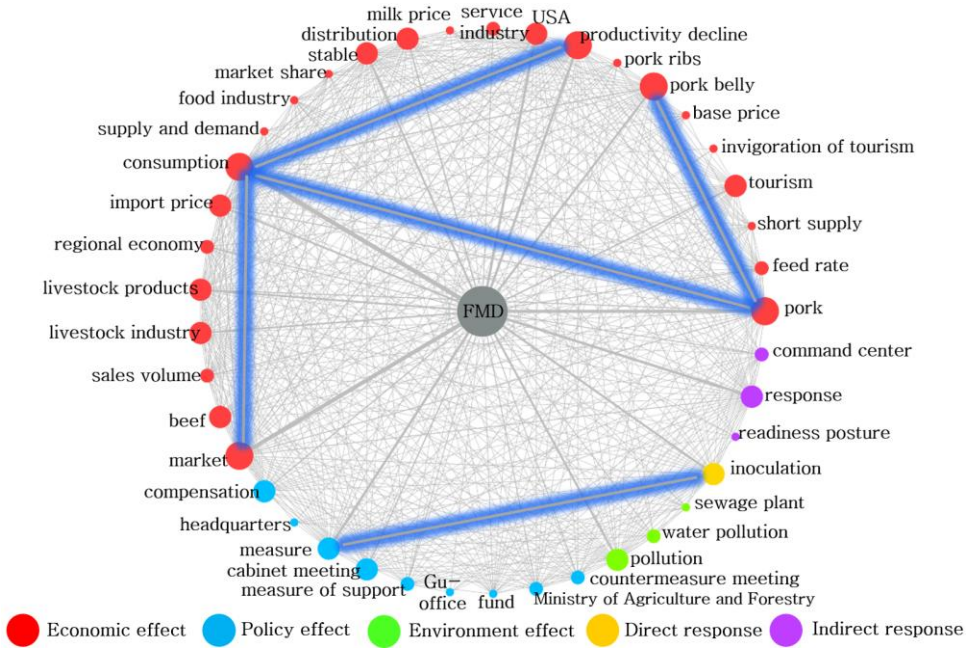


Figure 15. Keywords network in termination period

Table 9. Top 5 co- frequencies in termination period

No.	Co-keyword	Co-frequency
1	pork – pork belly	108
2	pork – consumption	90
3	measure – vaccination	83
4	consumption – market	82
5	productivity decline – consumption	72

*** Do not consider the co-frequency of 'FMD'-[keyword]**

4.3.4. Summary and Discussions

We conducted the analysis for relationships between keywords by constructing co-occurrence keywords network in each period. To sum up, in early period, in can be seen that keywords indicating direct/indirect responses were shown to be high and livestock industry were adversely affected due to FMD. In addition, through keywords such as 'prevention', it could be seen that efforts were made to prevent additional outbreaks of FMD. However, in serious period, many keywords indicating direct/indirect response are focused intensively such as 'disposals' and 'burying', rather than repercussion effects. Through these results, it can be seen that active responses were taken as the FMD outbreaks were serious. Finally, in termination period, co-frequencies of keywords pairs indicating most economic repercussion effects were shown to be high and some keywords related to indirect responses were emerged for the prevention of FMD in winter season. Meanwhile, policy and environmental repercussion effects did not focus as major issues in co-frequency perspective.

In this section, we constructed co-occurrence keywords network to analyze relationships between keywords. However, these analysis using it can obtain only one keyword pair (*e.g.* keyword-keyword). Therefore, in the next chapter, we will conduct the analysis comprehensively by using association

rule mining which generates rules among keywords. In particular, case studies will be conducted by focusing on rules of $\{action\} \rightarrow \{effect\}$ among diverse rules from ARM. According these rules, we can analyze the effects of diverse responses due to FMD outbreaks on repercussion effects.

4.4. Analysis Results using Association Rule Mining

4.4.1. Analysis in Early Period

As results of association rules, two or more issues appear complexly in the part of the consequence of the rules. In the early period, the most rules are represented in form of {{*response*} → {{*policy*} ∧ {*livestock industry*}}, as shown as **Table 10**. Unlike results of keywords network analysis, these results can be suitable for comprehensive analysis of social issues due to FMD.

From these rules, we can observe that many domestic animals died leading to serious adverse effects on the livestock industry due to responses actions such as ‘disposals’, ‘burying’ and ‘preventive measures’ (refer rules 1, 2, 3 and 5). Furthermore, through rule 4, it can be seen that countermeasure task forces were organized to minimize the economic adverse effects of FMD on the livestock industry and countermeasures and policies intended to such problems devised.

Table 10. Result of obtained rules in early period

No.	Association Rule			Sup.	Conf.
1	{disposals, burying}	→	{market share, Ministration of Agriculture and Forestry}	0.5	1.00
2	{command center, visitor, prevention, vehicle control}	→	{livestock products, market}	0.5	1.00
3	{non-proliferation}	→	{livestock products, beef, agriculture income}	0.5	1.00
4	{countermeasure force}	→	{livestock cooperatives, measure, policy, Rural Development Administration}	0.5	1.00
5	{command center, treat, check}	→	{livestock products, beef, agriculture income}	0.7	1.00

4.4.2. Analysis in Serious Period

In the results of analysis of association rules in the serious period, as shown in **Table 11**, most rules are represented in form of {{*response*} → {{*policy*} ∧ {*livestock industry*}}. These rules can be seen that efforts to minimize the economic adverse effects of FMD continued as rules 1, 2 and 3. Meanwhile, rules including environmental effects, which had not appear importantly in the OLAP analysis and network analysis, appeared as results. According to rules 4 and 5, we can observe that there are the water pollution problem in process of stamping out livestock such as blood and livestock wastewater. In addition, due to water pollution, even interregional conflicts appeared in relation to domestic animals burying regions.

Table 11. Result of obtained rules in serious period

No.	Association Rule			Sup.	Conf.
1	{emergency disease control}	→	{regional economy, countermeasure meeting, Ministration of Agriculture and Forestry}	0.33	1.00
2	{disposals, vaccination, emergency repair }	→	{regional economy, countermeasure meeting, Ministration of Agriculture and Forestry}	0.33	1.00
3	{command center, treat, emergency disease control }	→	{measure, Minister}	0.33	1.00
4	{burying, non-proliferation}	→	{regional conflict, pollution level, blood}	0.33	1.00
5	{control}	→	{management, regional conflict, leachate, livestock wastewater}	0.33	1.00

4.4.3. Analysis in Termination period

The association rules in termination period are represented in form of $\{\{indirect\ response\} \rightarrow \{\{other\ industry\} \wedge \{market\ economy\}\}$, as shown in **Table 12**. Rather than direct response, many indirect response appeared in the part of antecedents of the rules. In the rules 2, we can observe that there were negative effects in fields of tourism and food industries in the same manner as OLAP and network analysis. Meanwhile, issues for market economy were emerged such as 'demand and supply' and 'import price' in the part of consequences of the rules. These are issues arising from early delivery of livestock following the FMD outbreaks and it can be concluded that damage was expanded to consumers too due to the shortage of livestock product supply and price increases after termination.

Table 12. Result of obtained rules in termination period

No.	Association Rule			Sup.	Conf.
1	{command center, readiness posture, contact}	→	{livestock industry, consumption, supply and demand}	0.11	1.00
2	{treat, response}	→	{tourism, food industry}	0.11	1.00
3	{command center, prevention, treat, readiness posture, inoculation}	→	{regional economy, stable trend, distribution}	0.11	1.00
4	{command center, response}	→	{beef}	0.11	1.00
5	{treat, response}	→	{import price, pork}	0.22	1.00

4.4.4. Summary and Discussions

In this section, we performed the comprehensive analysis for causal relationships among keywords for each period by using association rule mining. Unlike the keyword network analysis, the results including the various issues were derived as rules. To sum up for each period, adverse effects of FMD on the livestock industry appeared and attempts to relieve such effects in terms of policies were shown in early period. In the serious period, environmental aspect related issues that had not become any issue in the OLAP analysis or network analysis appeared. Furthermore, as environmental problems incurred from burying and disposals, NIMBY phenomenon appeared that spread into interregional conflicts in relation to burying sites. In after termination period, many indirect response-related issues appeared so that it could be seen that the adverse effects appeared even in tourism and food industries. In addition, unlike the early and serious period, issues for supply and demand were derived as rules. These issues had been induced from early delivery of domestic animals following the outbreak of FMD and it can be concluded that damage was expanded to consumers too due to livestock price increases after termination of FMD.

4.5. Implementation

In this section, we present a 'livestock disease integrated control dashboard', which is implemented as prototype, to visualize the results of analysis and support to decision making. **Figure 16** comprehensively shows not only the results of OLAP analysis, network analysis and association rule mining but also the trend of outbreaks of FMD and related news to enable administrator to conduct more efficient and comprehensive analysis. This dashboard is expected to assist administrators in making more effective and strategic decisions.

Figure 16. Livestock disease integrated control system

5. CONCLUSIONS

In this thesis, we proposed a new design methodologies for the scrutinizing diverse social issues incurred from FMD in South Korea. Unlike the existing studies, online news, which immediately respond to diverse social issues and report only reliable information, were used. In order to confirm the feasibility and applicability of the proposed system, we conducted the analyzing online news for FMD in South Korea. First, the proposed system performed the multidimensional analysis by constructing keyword-cube model. Through this model, the issues, which cannot be found in public open data, were derived such as economic, policy and environmental effects. Moreover, whether or not appropriate responses were taken could be evaluated. Through network analysis and association rule mining, the relationships and causality among issues were analyzed with levels of abstraction, respectively. We believe that our results provide the valuable information that comprehensive issues from epidemic disease and administrator with making more effective decisions such as strategic devising policy.

Further, the refinement and testing of the proposed system are required to utilize this system into the advanced knowledge management system.

REFERENCES

[1] Mdetele, Daniel, et al. "Socio-economic Impact of Foot and Mouth Disease in Wildlife-Livestock Interface and Non-Interface of Tanzania." World 5.3 (2015): 31-35.
[2] Knight-Jones, T. J. D., and J. Rushton. "The economic impacts of foot and mouth disease–What are they, how big are they and where do they occur?." Preventive veterinary medicine 112.3 (2013): 161-173.
[3] Wekesa, S. N., et al. "Analysis of Recent Serotype O Foot-and-Mouth Disease Viruses from Livestock in Kenya: Evidence of Four Independently Evolving Lineages." Transboundary and emerging diseases 62.3 (2015): 305-314.
[4] Lycett, Samantha J., et al. "Role for migratory wild birds in the global spread of avian influenza H5N8." Science 354.6309 (2016): 213-217.
[5] Bradhurst, R. A., et al. "Improving the computational efficiency of an agent-based spatiotemporal model of livestock disease spread and control." Environmental Modelling & Software 77 (2016): 1-12.
[6] Muroga, Norihiko, et al. "The 2010 foot-and-mouth disease epidemic in Japan." Journal of Veterinary Medical Science 74.4 (2012): 399-404.
[7] Xu, Zhenshun, et al. "Propagation Routes Analysis of HPAI Outbreaks using Sequential Pattern Mining." Proceedings of The fourth International Conference on Information Science and Cloud Computing (ISCC2015). 18-19
[8] Lee, Hyung-jin, et al. "Prediction of the spread of highly pathogenic avian influenza using a multifactor network: Part 2–Comprehensive network analysis with direct/indirect infection route." Biosystems engineering 118 (2014): 115-127.
[9] Rushton, Jonathan. The economics of animal health and production. Cabi, 2009.
[10] Ferrari, G., et al. "Foot-and-Mouth Disease and Its Effect on Milk Yield: An Economic Analysis on Livestock Holders in Pakistan." Transboundary and emerging diseases 61.6 (2014): e52-e59.
[11] Carpenter, Tim E., et al. "Epidemic and economic impacts of delayed

detection of foot-and-mouth disease: a case study of a simulated outbreak in California." Journal of Veterinary Diagnostic Investigation 23.1 (2011): 26-33.
[12] Hou, Lei, et al. "NewsMiner: Multifaceted news analysis for event search."Knowledge-Based Systems 76 (2015): 17-29.
[13] Ghosh, Saurav, et al. "Temporal Topic Modeling to Assess Associations between News Trends and Infectious Disease Outbreaks." arXiv preprint arXiv:1606.00411 (2016).
[14] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.
[15] Zhu, Jiahui, et al. "Coherent topic hierarchy: A strategy for topic evolutionary analysis on microblog feeds." International Conference on Web-Age Information Management. Springer International Publishing, 2015.
[16] Dou, Wenwen, et al. "Hierarchicaltopics: Visually exploring large text collections using topic hierarchies." IEEE Transactions on Visualization and Computer Graphics 19.12 (2013): 2002-2011.
[17] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.
[18] Kimball, Ralph, and Margy Ross. "The data warehouse toolkit: the complete guide to dimensional modelling." Nachdr.]. New York [ua]: Wiley (2002): 1-447.
[19] Loudcher, Sabine, et al. "Combining OLAP and information networks for bibliographic data analysis: a survey." Scientometrics 103.2 (2015): 471-487.
[20] Zhang, Duo, ChengXiang Zhai, and Jiawei Han. "MiTexCube: MicroTextCluster Cube for Online Analysis of Text Cells." CIDU. 2011.
[21] Liu, Xiong, et al. "SocialCube: A Text Cube Framework for Analyzing Social Media Data." Social Informatics (SocialInformatics), 2012 International Conference on. IEEE, 2012.
[22] Zhang, Duo, Chengxiang Zhai, and Jiawei Han. "Topic Cube: Topic Modeling for OLAP on Multidimensional Text Databases." SDM. Vol. 9. 2009.
[23] Yang, Dong Hoon, et al. "Association rule mining and network analysis in oriental medicine." PLoS One 8.3 (2013): e59241.
[24] Dorjee, S., et al. "Network analysis of swine shipments in Ontario, Canada, to support disease spread modelling and risk-based disease management." Preventive veterinary medicine 112.1 (2013): 118-127.

[25] Hu, Jiming, and Yin Zhang. "Research patterns and trends of Recommendation System in China using co-word analysis." Information Processing & Management 51.4 (2015): 329-339.
[26] Li, Huajiao, et al. "Words analysis of online Chinese news headlines about trending events: A complex network perspective." PloS one 10.3 (2015): e0122174.
[27] Lu, Xin, and Christa Brelsford. "Network structure and community evolution on twitter: human behavior change in response to the 2011 Japanese earthquake and tsunami." Scientific reports 4 (2014): 6773.
[28] Soysal, Ömer M. "Association rule mining with mostly associated sequential patterns." Expert Systems with applications 42.5 (2015): 2582-2592.
[29] Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey. "Business Intelligence and Analytics: From Big Data to Big Impact." MIS quarterly 36.4 (2012): 1165-1188.
[30] Kamsu-Foguem, Bernard, Fabien Rigal, and Félix Mauget. "Mining association rules for the quality improvement of the production process." Expert systems with applications 40.4 (2013): 1034-1045.
[31] Lin, Shiang-Lin, et al. "ANALYZING MEDICAL TRANSACTION DATA BY USING ASSOCIATION RULE MINING WITH MULTIPLE MINIMUM SUPPORTS." (2016).
[32] Nahar, Jesmin, et al. "Association rule mining to detect factors which contribute to heart disease in males and females." Expert Systems with Applications 40.4 (2013): 1086-1093.
[33] Rose, J. Dafni. "AN EFFICIENT ASSOCIATION RULE BASED HIERARCHICAL ALGORITHM FOR TEXT CLUSTERING." Int J Adv Engg Tech/Vol. VII/Issue I/Jan.-March 751 (2016): 753.
[34] Hurtado, Jose L., Ankur Agarwal, and Xingquan Zhu. "Topic discovery and future trend forecasting for texts." Journal of Big Data 3.1 (2016): 1.
[35] http://www.kahis.go.kr/ [Accessed: May 31. 2016]
[36] Jeon, H. "KoNLP: Korean NLP package." R package version 0.76 8 (2012).
[37] Hornik, Kurt, and Bettina Grün. "topicmodels: An R package for fitting topic models." Journal of Statistical Software 40.13 (2011): 1-30.

[38] Jones, Isaac, Lei Tang, and Huan Liu. "Community Discovery in Multi-Mode Networks." User Community Discovery. Springer International Publishing, 2015. 55-74.
[39] Latapy, Matthieu, Clémence Magnien, and Nathalie Del Vecchio. "Basic notions for the analysis of large two-mode networks." Social networks 30.1 (2008): 31-48.
[40] Latapy, M., Magnien, C., Del Vecchio, N., 2008. Basic notions for the analysis of large two-mode networks. Social Networks 30(1), 31-48.
[41] Csardi, Maintainer Gabor. "Package 'igraph'." The Comprehensive R Archive Network. See http://cran. r-project. org/web/packages/igraph/igraph. pdf (2015).
[42] Hahsler, Michael, et al. "Package 'arules'." (2014).
[43] Havre, Susan, et al. "Themeriver: Visualizing thematic changes in large document collections." IEEE transactions on visualization and computer graphics 8.1 (2002): 9-20.