Approach for visualizing research trends: three-dimensional visualization of documents and analysis of relative vitalization

Sang-Jun Yea, Chul Kim

Information Development & Management Group,
Korea Institute of Oriental Medicine, Yuseong-gu Daejeon, 305-811, Korea

ABSTRACT

This paper proposes an approach for visualizing research trends using theme maps and extra information. The proposed algorithm includes the following steps. First, text mining is used to construct a vector space of keywords. Second, correspondence analysis is employed to reduce high-dimensionality and to express relationships between documents and keywords. Third, kernel density estimation is applied in order to generate three-dimensional data that can show the concentration of the set of documents. Fourth, a cartographical concept is adapted for visualizing research trends. Finally, relative vitalization information is provided for more accurate research trend analysis. The algorithm of the proposed approach is tested using papers about Traditional Korean Medicine.

Key words: *Research Trend Visualization, Information Visualization, Vitalization Analysis, Theme Map.*

1. INTRODUCTION

People live in three-dimensional space and recognize objects in the real world. And people analyze and integrate the information to use the objects around them. A map is invented to utilize the people's spatial perception ability [1], [2]. In general, map is used to know where we are and how to get to the destination. Also this concept can be used to express the underlying structure in the documents set. It is called as 'Theme Map' which is expressed by three-dimensional landscapes [2]-[4]. And the map will convey relational information between documents. If theme map is constructed in three-dimension, then it requires the changing of view point and the integration of information to analyze the relationship among documents [2], [5]. This extra effort makes hard to get insight for documents' relation from theme map, it arises needs for alternative concept, for example bird eye view [1].

A contour map is made to provide the bird eye view for the topography of three-dimensional landscape [6], [7]. And we can use the contour map metaphor to grasp the documents' relationship. A contour map metaphor is made good use in the IT field to visualize the mutual relation - web structure, software procedure, document relationship and so on [2]. To construct a contour map, we must extract the information which can best represent documents and can express the underlying relationship. In general, contour map is composed of lines and colors which are called as contour lines and depth colors respectively [2], [7]. If any two documents are similar, then they placed much closer than any others on the theme map.

Peaks appear on the map where there is a high concentration of documents. The distance between valleys and ridges shows how closely the themes are related. To construct and visualize the contour map, 3 steps algorithm - information analysis, dimension reduction and map visualization are needed [2], [6], [7].

The challenge of this research aims to give high quality research trend analysis tools through creating theme map and extra information. This research adopts Traditional Korean Medicine (TKM) papers to show the algorithms and procedures suggested in this study.

2. RELATED TECHNOLOGIES

2.1 Information analysis

To visualize the information on the theme map, information analysis is needed which is composed of indexing and analysis process [1]. Indexing process is to extract the information from unstructured documents. It is composed of keyword extraction stage using natural language processing and vector space construction from the extracted keywords. Automatic indexing is a familiar algorithm to show the corpus space of each document. And a natural language processing can represent document's content well. The part-of-speech-tagging is the basic technology in the noun processing algorithm. And this method is more accurate than the indexing based algorithms [8].

Analysis process is to discover the extracted information from indexing stage. In general, it uses the classification and clustering methods. The classification algorithm is assigning the document to the predefined groups. But the clustering

method can create variable groups based document's similarity. Widely used classification methods include the naive Bayesian method, k-nearest neighbor and network models. And commonly used clustering algorithm is self-organizing map (SOM) which produces a two-dimensional grid representation for N-dimensional features [7]. Other popular clustering algorithms include multidimensional scaling, the k-nearest neighbor method, and the K-means algorithm [9].

2.2 Dimension reduction

Each document is composed of keyword vector and each keyword means one dimension. So the vector space of documents' set stores information in the high-dimension. Dimension reduction is needed to visualize the relationship in the high-dimensional vector space. Commonly two-dimensional or three-dimensional visualization is preferred. This process always arise information distortion problem. Multidimensional Scaling (MDS), Self Organizing Map (SOM), Principal Component Analysis (PCA), Correspondence Analysis (CA) is commonly used [8], [10]. MDS is a set of related statistical techniques often used in information visualization to show the relationship and structure of data. MDS is commonly used to show the ordination results. MDS algorithm deals with the similarity matrix and it locate each item in n dimensional space – typically two or three dimension [10].

SOM is a kind of unsupervised machine learning using artificial intelligence methods. SOM produce low-dimensional representation of the input space called a map. SOM is widely used to visualize high-dimensional data in low-dimensional space. The SOM model was first described as an artificial neural network by the Finnish professor Teuvo Kohonen [7].
PCA adopts a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components [11]. The first principal component has the most information about the original data, and each succeeding components has the left information as possible. If the set is composed of high-dimensional data which is hard to view, then PCA can supplies the user with a lower-dimensional picture, a "shadow" of this object. And this selected dimension (principal components) is the most informative viewpoint.

CA was developed by Jean-Paul Benzécri and is conceptually similar to PCA, but scales the data so that rows and columns are treated equivalently [12]. CA commonly deals with the contingency table or matrix, and decomposes the table into orthogonal factors using chi-square method. CA can produce the relations between rows and columns. CA has the distinct advantage over other dimension reduction algorithm. It is called duality which can reveal non-linear relationships between rows and columns of a multidimensional contingency table.

2.3 Map visualization

There are seven kind of information visualization methods as 1-D, 2-D, 3-D, multidimensional, tree, network and temporal [13]. Among seven categories, we reviewed the

multidimensional methods which have capacity of representing map data. The multidimensional approach represents information as multidimensional objects and projects them into a three-dimensional or a two-dimensional space. The SPIRE system is belong to this category [14].

The ThemeScape is the documents analysis tool using theme map in the SPIRE system which is implemented to help solve the problem of information overload [6]. And There are several graphical tools such as MATLAB[15], Mathematica[16], ChartFX[17], TeeChart[18], and R package[19] etc. which can represent three dimensional data as contour map.
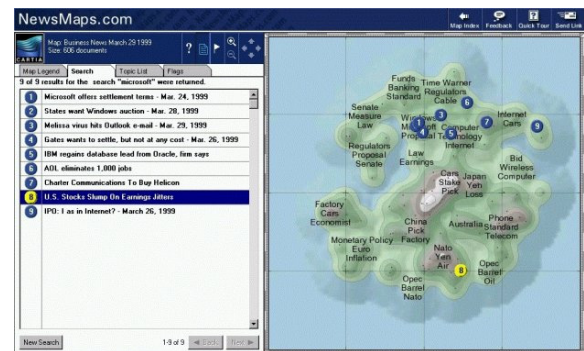

Fig. 1. An example of ThemeScape

3. ALGORITHM

This section explains the algorithm to make theme map using contour map as stated before. The algorithm is composed of the following seven steps:

- ✓ Vector space formation
- ✓ Correspondence analysis
- ✓ Kernel density estimation
- ✓ Local maxima detection
- ✓ K-nearest neighbor detection
- ✓ Contour map generation
- ✓ Research vitalization analysis

In implementation, the output of former step is used as the input of latter step. And the algorithm is executed sequentially. First, the vector space of matrix form is created from the extracted keywords which came from documents. Next, correspondence analysis is applied to the former vector space to get the two-dimensional coordinates of documents. And density probability estimation transforms it to get the smoothed curved surface. To locate the peak, local maxima detection and k-nearest neighbor detection are used. And the data are visualized as contour map by the R package. Finally the former data are analyzed to get the vitalization information for the documents.

Step one: vector space formation

The analysis of natural language (unstructured) document provides a characterization on content. This process is composed of extraction of essential descriptor and representation of these features. Text features are typically one of three general types, though any number of variations and

hybrids are possible [4]. The first type is frequency-based on first order statistics. The count of key words in a document is measured and used as feature vector. The second type is on higher order statistics captured by Bayesian or neural nets. The third type is semantic in nature. This approach utilizes word corpus and the knowledge of language to extract the semantic in the text.

We used the keyword frequency method to create vector space like Figure 2. To extract the experts' knowledge, the keywords (Keyword 1, Keyword 2, …, Keyword n) are selected from the term dictionary [20]. In the Figure 2, the keyword's counts are express as C_{mn} for each document (Document 1, Document 2, …, Document m)

Fig. 2. The matrix of documents-keyword vector space

Step two: correspondence analysis

CA is a multivariate descriptive method based on a data matrix with non-negative elements and related to principal component analysis (PCA). In general CA is applied to categorical data, but it can be applicable for any kind of data. CA has duality characteristics which came from the scaling the data, and deals rows and columns equivalently. Thus we can induce the relation of not only within row data and within column data but also between row data and column data.

The CA solution was shown to be neatly encapsulated in the singular-value decomposition (SVD) of a suitably transformed matrix [12]. To summarize the theory, first divide the I × J data matrix, denoted by N, by its grand total n to obtain the so-called correspondence matrix $P = N/n$. Let the row and column marginal totals of P be the vectors r and c respectively, that is the vectors of row and column masses, and D_r and D_c be the diagonal matrices of these matrices. The computational algorithm to obtain coordinates of the row and column profiles with respect to principal axes, using the SVD, is as follows:

- Calculate the matrix of standardized residuals: $S = D_r^{-\frac{1}{2}}(P - rc^T)D_c^{-\frac{1}{2}}$
- Calculate the SVD: $S = UD_\alpha V^T$ where $U^T U = V^T V = I$
- Principal coordinates of rows: $F = D_r^{-\frac{1}{2}}UD_\alpha$
- Principal coordinates of columns: $G = D_c^{-\frac{1}{2}}VD_\alpha$
- Standard coordinates of rows: $X = D_r^{-\frac{1}{2}}U$
- Standard coordinates of columns: $Y = D_c^{-\frac{1}{2}}V$

Finally we can get the two dimensional coordinates of each document (row) and keyword (column) from X and Y.

Step three: kernel density estimation

In mathematics, extrapolation is the algorithm to construct new point from set of known points. It is similar to the interpolation process which constructs new inner point between known points. There are several kinds of extrapolation which is widely applied in computation. The linear extrapolation and polynomial extrapolation is commonly used, but the result which is constructed from the method is poor. And Conic extrapolation and French curve extrapolation generate better results than previously mentioned algorithms, but theses uses more points to construct new point. Thus it consumes more computational power [21].

In statistics, kernel density estimation (KDE) is a non-parametric way of estimating the probability density function of a random variable. Given some data about a sample of a population, kernel density estimation makes it possible to extrapolate the data to the entire population [22]. It means that we can construct three dimensional coordinate from the results of CA through applying KDE. If $x_1, x_2, \ldots, x_n \sim f$ are independent and sample of random variables which follow same probability distribution, then kernel density approximation is

$$f_h(x) = \frac{1}{Nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \text{ where } K\left(\frac{x-x_i}{h}\right) = \frac{1}{\sqrt{2\pi}}e^{-\frac{(x-x_i)^2}{2h^2}}$$

Equation 1. Kernel density approximation function

At this point, K is some kernel and h is smoothing parameter called bandwidth.

Step four: local maxima detection

Local maxima detection and local minima detection are used to detect points which have outstanding characteristics or features. It is commonly applied in image processing of computer vision fields. There are several kinds of feature detection algorithm. Corner detection and blob detection algorithms are commonly used in early ages. Image features which must be well-defined and stable are extracted with corner detection. And blob detection is very similar to the local maxima detection in using blob descriptor. Recently interest point detection is more common terminology in image processing. An interest point is a point in the image which in general can be characterized as follows: [23]

- It has a clear, preferably mathematically well-founded, definition.
- It has a well-defined position in image space.
- The local image structure around the interest point is rich in terms of local information contents.

We used local maxima detection to locate the peak of three dimension data constructed from KDE. If the x satisfies $f(x) \geq f(x')$ for all x' located around x, then we call $f(x)$ as local maxima. That is, if the x satisfies $\|x - x'\| < \varepsilon$ for some $\varepsilon > 0$, then all x must satisfy previous formula. Because we deal with three dimension data, we have to find all (x, y) positions which satisfy $f(x, y) \geq f(x', y')$. Because these detected peaks are the locally highest points of the three dimension data, these positions are where the density of documents is high.

Step five: k-nearest neighbor detection

Nearest neighbor detection (NND), also known as similarity detection or closest point detection, is an optimization problem for finding closest points in metric spaces [24]. There are numerous variants of the NND problem and the two most well-known are the k-nearest neighbor detection (k_{NN}) and the ε-approximate nearest neighbor detection. The k_{NN} detection identifies the top k nearest neighbors to the point. We can compute the distance between two points using some Euclidean distance functions like below.

$$d(x,y) = \sqrt{(x - x')^2 + (y - y')^2}$$

where (x, y) is peak position
and (x', y') is keyword position

Equation 2. Euclidean distance function

For all peak points and all keyword positions, the Euclidean distance are measured. And k_{NN} identifies the top k nearest neighbor keyword positions to the peak which is detected by KDE. This means that the most common k keywords are displayed in the peak points.

Step six: contour map visualization

A contour line (also isoline, isogram or isarithm) of a function of two variables is a curve along which the function has a constant value [25]. In cartography, contour map is illustrated with contour lines which are set of lines. A Contour line is the joined points of equal height. Contour lines show the hills, valleys and peaks which can represent the steepness of slopes, show the distance of peaks and so on. And contour level is the successive contour lines which are represented by the interpolated colors. It helps people understand the contour map. In general, contour plot is a graphic representation of the relationships among three numeric variables (x, y and z) in two dimensions using contour map metaphor.

There are several graphical tools such as MATLAB [15], Mathematica [16], ChartFX [17], TeeChart [18], and R package [19] etc. which can represent three dimension data as contour map. We choose the R package as graphical tools because this software provides java connection which can develop the web based program. Besides this function it has plenty of statistical libraries to help the numeric and statistical calculation.

Step seven: research vitalization analysis

After the construction of theme map, we can analyze the height of the peaks. Correspondence analysis can represent the relation between the document and keyword. If we analyze the peaks' height, we can get the relative vitalization of the documents. Because the absolute vitalization of documents requires more statistical information, we only know about the relative vitalization information. It is divided as many stages as the user want to. In this study, we used three-step stages (low, mid, high) relative vitalization. Below is the detailed process of vitalization analysis:

- Calculate the difference between the highest peak and ground
- Divide the height into three and get the each height
- Create the (x,y) positions classified as each stage

- Get the vitalization of the documents and keywords using each position

4. AN EXAMPLE OF IMPLEMENTATION: THE CASE OF TKM PAPERS

4.1 Test bed: a TKM paper collection

We used Korea Institute of Oriental Medicine's Oriental Advanced Searching Integrated System (OASIS) to test the developed algorithm. OASIS is the largest database service system in Korea in the field of TKM [26].

Table 1. Data fileds serviced by OASIS

Data Field	Data Type
Title (Korean and English)	String
Authors (Korean and English)	String
Keyword (Korean and English)	String
Abstract (Korean and English)	String
Category	String
Organ	String
Journal	String
Issue	Date
Volume	String
ISSN	String
Full text	PDF File link

Table 1 shows the provided information by the OASIS. And we extracted title, keyword, and abstract from the information. To construct the experts' term dictionary, we used 22,334 TKM terms from the TKM Korean-English dictionary [20]. And we used 11,160 medical terms form then Korean Medical Library Engine [27].

4.2 Theme map creation

We selected 252 papers which searched using 'ginseng' keyword from OASIS. First the system calculates each keyword's frequencies from searched papers. And the system cut off the keyword list using threshold to get the high quality visualization. Using the keyword list, it generates the paper-keyword matrix which is used as vector space and is fed as the CA input. CA was applied for the paper-keyword matrix and it generates two-dimensional data for the papers and keywords. The last of algorithm are applied as stated before. And finally we get the theme map in Figure 3, but the red color letter is not generated by the system. It is marked to provide the analysis example.
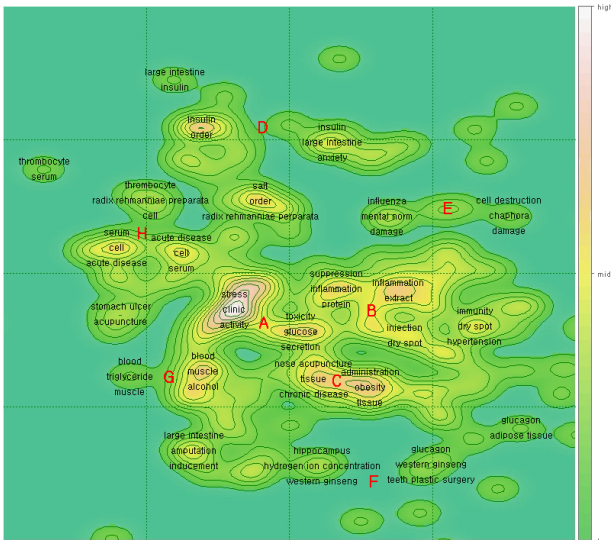
Fig. 3. Theme map generated by ginseng related papers

4.3 Research trend analysis

The implemented system can generate research vitalization information based on three-step stages. It assigns each paper and keyword into one of the stages. The generated keyword list is like Table 2. High vitalization contains 33 keywords and mid vitalization contains 31 keywords. Finally low vitalization contains 22 keywords. Among 252 papers, 131 are assigned as high vitalization stage, 53 are mid stage, and 68 are low stage.

Table 2. Keyword list for each vitalization stage

Vital-ization	Keyword
high	dry spot, hypertension, obesity, muscle, pathology, diet, toxicity, large intestine, chronic disease, immunity, cultivation, secretion, nose acupuncture, lack of oxygen, stress, cytokine, gall bladder, drug, inflammation, temperature, stomach, clinic, tissue, injection, addiction, order, weight, extract, acupuncture, cholesterol, administration, glucose, activation
mid	cold, solid, stomach ulcer, glucagon, basis, protein, femur, deoxyribonucleic acid, expression, valves, rat, western ginseng, cell, cell toxicity, cell destruction, radix rehmanniae preparata, bile, alcohol, herb, suppression, salt, gene, ginseng, insulin, interleukin, amputation, mental norm, hippocampus, blood, serum, enzyme
low	acute disease, metabolism, comparison, animal, intestine, immunity reaction, spot, anxiety, damage, hydrogen ion concentration, diet, suppression effect, inducement, sarcoma, chapora, tumor, case, symptom, teeth, triglyceride, platelet, patient

Now we can analyze the research trend more accurately using theme map and vitalization keyword list. Figure 3 can be analyzed like below. This analysis is done only by the TKM experts.

A: Researches about ginseng's effect for stress
B: Researches about ginseng's effect for inflammation suppression
C: Researches about ginseng's tissue
D: Researches about combined administration ginseng and insulin
E: Researches about chaphora ginseng's effect for cell destruction
F: Researches about western ginseng's effect
G: Researches about ginseng specific component's effect for blood and muscle
H: Researches about ginseng's effect for blood and cell

4.4 Performance evaluation

In order to provide a performance evaluation and further evidence of the scalability of our methodology. Performance evaluation environment is set as web server and database servers are separated.

- Web server : Intel Pentium D 3.40GHz, 2.50GB RAM, Windows XP SP 3, Tomcat 5.5
- Database server : 2 Xeon 3.0GHz, 2GB, Solaris 9, Oracle 9i

We used System.currentTimeMillis() function to measure the real operational time in the evaluation environment. And the result is like Table 3. We averaged 10 evaluation results as summarized in Table 3. Vector space formation, correspondence analysis, kernel density estimation, and local maxima estimation algorithms were executed within 100ms. And visualization was executed about 800ms and vitalization analysis algorithm almost took no times. The total time for information visualization was 1053ms

Table 3. Performance evaluation result

Processing	Time(millisecond)
Vector space formation	21
Correspondence analysis	78
Kernel density estimation	93
Local maxima estimation	78
Visualization	782
Vitalization analysis	1

5. DISCUSSION & CONCLUSIONS

In this study, we showed the new theme map visualization algorithm based on the keyword. Also we suggested new analysis algorithm to get the relative vitalization. Finally we illustrated that it is possible to analyze the research trend through theme map and vitalization information. Using keyword extracted from documents has several advantages in information retrieval, visualization, and analysis than using references of documents. Keyword based theme map analysis

system can maximize the systemized portion due to reduce extra experts' knowledge.

If each document is not represented by a high dimensional vector whose components indicate that document's discriminating words and how those words are connected to all other topics of interest that span and describe the document collection, then visualization tools may cause the bottleneck and inaccurate placement that occurs when a landscape visualization like a ThemeScape is based on an intervening document projection to the two-dimensional plane as occurred with Galaxies [6]. In this paper we adopted correspondence analysis whose row and column geometries have similar interpretation, so we avoided the above mentioned problems.

In the map, height is created by the proportion of the papers' density. So we can get relative information for the vitalization of research using the calculation of height. It is the difference between ground and the highest peak on the map. The height is divided into trisection, and each two-dimensional position is calculated. All documents' position and all keywords' positions are compared to the former locations. The developed system provides document list and keyword list related to each stage. The researcher analyzes the research trend with theme map and vitalization information.

It spends much time and experts' knowledge to grasp the research trend and to decide the new research area from the paper analysis manually. In this study, we showed that it spends less time and efforts when we use the theme map and extra information, for example vitalization information. We used TKM paper database to example developed algorithms. But the same technique can applied to the patent, project report, and research note etc.

By its nature, this study is an exploratory one, and needs more extension and elaboration in terms of methodology and application. First, it is need to analyze the absolute vitalization stage. It will be impossible to get the goal if we use only theme map's information. It means that extra information is need. Second, more severe test process must be performed to get trust for this result. Test for research materials in the various fields must be measured to show that this system is useful. And the comparisons are need with the experts' analysis results. Third, CA has the advantage to get the relationship between documents and keywords. But like PCA, MDS, SOM, and etc. there is weakness called information loss due to the dimension reduction. Thus new algorithm is needed to reduce this information loss.

ACKNOWLEDGEMENTS

REFERENCES

[1] ASIS&T, *Annual review of information science and technology 39th, Information Today*, Medford NJ, 2005.

[2] W. Chung, H. Chen, and J. F. Nunamaker Jr, "Business Intelligence Explorer: A Knowledge Map Framework for Discovering Business Intelligence on the Web," Proceedings of the 36th Hawaii International Conference on System Science, 2003.

[3] E. C. Engelsman and A. F. J. van Raan, "A patent-based cartography of technology," Research Policy, vol. 23, 1994, pp. 1-26.

[4] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, "Visualizing the Non-Visual: Spatial analysis and interaction with information from text documents," Proceedings of the 1995 IEEE Symposium on Information Visualization, Atlanta, Georgia, 1995, pp. 51-58.

[5] P. G. Anick and S. Vaithyanathan, "Exploiting clustering and phrases for context based information retrieval," Proceedings of the ACM SIGIR Annual International Conference on Research and Development in Information Retrieval, 1997, pp. 314-323.

[6] A. James, "Wise The Ecological Approach to Text Visualization," Journal of the American society for information science, vol. 50, no. 13, 1999, pp. 1224-1233.

[7] S. Kaski, "Data exploration using self-organizing maps," Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering, Series 82, 1995, p. 57.

[8] C. Salton, *Automatic text processing,* Addison-Wesley, Reading MA, 1989.

[9] Y. Y. Yang, L. Akers, T. Kose, and C. B. Yang, "Text mining and visualization tools – Impressions of emerging capabilities," World Patent Information, vol. 30, no. 4, 2008, pp. 280-293.

[10] J. B. Kruskal and M. Wish, *Multidimensional Scaling, Sage University Paper series on Quantitative Application in the Social Sciences*, Sage Publications, Beverly Hills and London, 1978.

[11] Y. Theodorou, C. Drossosb, and P. Alevizos, "Correspondence analysis with fuzzy data: the fuzzy eigenvalue problem," Fuzzy Sets and Systems, vol. 158, 2007, pp. 704-721.

[12] Greenacre, *Michael Theory and Applications of Correspondence Analysis*, Academic Press, London, 1983.

[13] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualization," Proceedings of IEEE Workshop on Visual Language, 1996, pp. 336-343.

[14] K. W. Boyack, B. N. Wylie, and G. S. Davidson, "Domain visualization using VxInsight for science and technology management," Journal of the American Society for Information Science and Technology, vol. 53, no. 9, 2002, pp. 764-774.

[15] Mathwork website. Available at: < http://www.mathworks.co.kr>

[16] Wolfram website. Available at: http://wolfram.com

[17] SoftwareFx website. Available at: < http://www.softwarefx.com>

[18] Teechart website. Available at: < http://www.teechart.com>

[19] R-project website. Available at: < http://www.r-project.org>

[20] Dictionary publishing committee, *Korean-English Dictionary of Oriental Medicine*, Jimundang, 2004.

[21] C. Brezinski and Z. M. Redivo, *Extrapolation Methods: Theory and Practice*, North-Holland, 1991.

[22] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, Springer Texts in Statistics, 2005.

[23] T. Lindeberg, "Detecting Salient Blob-Like Image Structures and Their Scales with a Scale-Space Primal Sketch: A Method for Focus-of-Attention," International Journal of Computer Vision, vol. 11, no. 3, 1993, pp. 283-318.

[24] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is nearest neighbor meaningful?," In Proceedings of the 7th ICDT, 1999.

[25] C. Richard, H. Robbins, and I. Stewart, *What Is Mathematics: An Elementary Approach to Ideas and Methods*, Oxford University Press, New York, 1996.

[26] OASIS website. Available at: < http://oasis.kiom.re.kr>

[27] Korean Medical Library Engine website. Available at: <http://www.kmle.co.kr>

Yea Sang-Jun
He received the B.S., M.S in computer science from KAIST, Korea in 2002, 2004 respectively. From 2008, he has been with the Information Development & Management Group, Korea Institute of Oriental Medicine.

Kim Chul
He received the B.S., M.S in industrial engineering from KAIST, Korea in 1998, 2000 respectively and also received Ph.D. in oriental medicine informatics from Wonkwang university, Korea in 2009. From 2006, he has been with the Information Development & Management Group, Korea Institute of Oriental Medicine.