divide the group. In this study, we define VAS as distribution and develop a new clustering method to this dataset directly using Symbolic data analysis.

2. Symbolic data analysis (SDA)

Conventional data analysis usually can handle scalars, vectors and matrices. However, lately, some datasets have grown beyond the framework of conventional data analysis. Most statistical methods do not have sufficient power to analyze these datasets.

Symbolic data analysis (SDA) proposed by Diday [2],[4] is an approach for analyzing new types of datasets. "Symbolic data" consist of a *concept* that is described by intervals, distributions, etc. as well as by numerical values. The use of SDA enriches data description, and it can handle highly complex datasets. This implies that complex data can be formally handled in the framework of SDA. However, most SDA works have dealt with only intervals as the descriptions and are very few studies based on this simple idea. The case that *concept* is described by intervals is simple, but ignores detailed information in the intervals. We propose distribution-valued data to describe the *concept.*

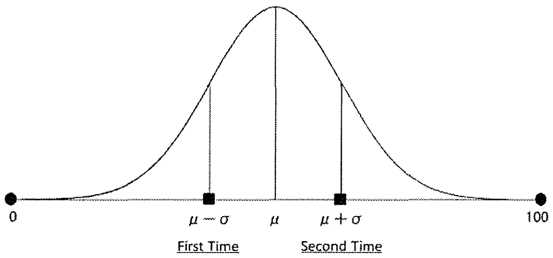

Figure 2Transform the Visual Analogue Scale to pre-PD

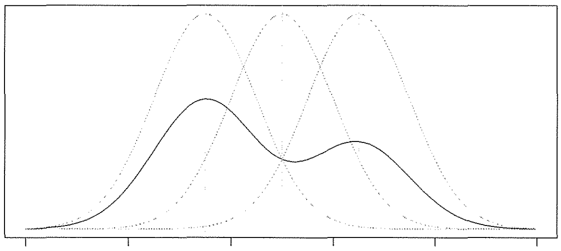

Figure 1 Transform the Visual Analogue Scale to PD: first time=35 second time=65

3. The Visual Analogue Scale (VAS)

A VAS consists of a line on a page with clearly defined end points, and normally a clearly identified scale between the two end points. For guidance, the phrase "no pain" and "worst imaginable pain" are placed at the both side of the line, respectively. Minimum value 0 of the VAS means "no pain" and maximum value 100 means "worst imaginable pain". These scales are of most value when looking at change within patients, and are of less value for comparing across a group of patients because patient have a different sense of pain. It could be argued that a VAS is trying to produce interval/ratio data out of subjective values that are at best ordinal. Thus, some caution is required in handling such data. Many researchers prefer to use a method of analysis that is based on the rank ordering of scores rather than their exact values, to avoid reading too much into the precise VAS score.

4. Transform the Visual Analogue Scale to Distribution-Valued Data

We transform the VAS to distribution-valued data. We suggest that sense of pain is described by mixture normal distribution and call it "pain distribution (PD)" Let VAS score of patient's first time be x1 and second time be x2. To define PD, we set pre-PD. We define the middle point of x_1 and x_2 as mean of each patient μ, and $(\mu - x_1)^2 = (\mu - x_2)^2$ as variance. We describe pre-PD as $N(\mu, \sigma^2)$ (Figure 1)..

Next, We set score of patient's first time x1 be mean of pre-PD as new distribution of first time (pre-PD1). New distribution of second time is defined in a similar way(pre-PD2). By combining pre-PD1 and pre-PD2, we get mixture distribution, where set mixture weight for pre-PD1 as 0.6 and for pre-PD2 as 0.4. It is, finally, PD. Figure 2 shows the case that x_1 is 35 and x_2 is 65. In case that the number of VAS score is *d,* PD is *d*-dimensional normal distribution. In this case, a diagonal matrix is used as a variance-covariance matrix of d-dimensional normal distribution.

5. Hierarchical Clustering for PD

Cluster analysis groups data objects only on the bases of information found in the data that describes the objects and their relationships. The goal is that the objects within a group should be similar (or related) to one another and different from the objects in other groups.

In this section, we propose a hierarchical clustering for distribution-valued data, especially for PD.

5.1. The Clustering Algorithm

We extend the idea of a hierarchical clustering in the framework of conventional data analysis. Let *n* be the number of PD and *K* be the number of cluster.

<Step1> Begin with *K* clusters, each containing only a single PD, *K* = *n*. Calculate distance between PD.

<Step2> Search the minimum distance in *K* clusters. Let the pair the selected clusters. Combine PDs into a new cluster, It is described by mixture distribution of the member, where mixture weight is equal. Let *K* be *K*-1. If *K* > 1, go to Step3, otherwise Step4.

<Step3> Calculate the distance between new cluster and other cluster, and go back to Step2.

<Step4> Draw the dendrogram.

Kullback-Leibler divergence is the natural way to define a distance measure between probability distributions [5], but not symmetry. We would like to use the symmetric Kullback-Leibler (symmetric KL) divergence as distance between *concepts*. The symmetric KL-divergence between two distributions s_1 and s_2 is

$$\begin{aligned} D\big(s_1(x), s_2(x)\big) &= D\big(s_1(x) \parallel s_2(x)\big) + D\big(s_2(x) \parallel s_1(x)\big) \\ &= \int_{-\infty}^{\infty} s_1(x) \log \frac{s_1(x)}{s_2(x)} dx + \int_{-\infty}^{\infty} s_2(x) \log \frac{s_2(x)}{s_1(x)} dx \end{aligned} \qquad (1)$$

where $D\big(s_1(x) \parallel s_2(x)\big)$ is KL divergence from s_1 to s_2 and $D\big(s_2(x) \parallel s_1(x)\big)$ is one from s_2 to s_1.

5.2. Distance Between PDs

In section 5.1, we use symmetric KL-divergence as distance between PDs. It is symmetric KL-divergence between Gaussian mixture distributions. However, it cannot be analytically computed. We can use, instead, Monte-Carlo simulations to approximate the symmetric KL-divergence. The drawback of the Monte-Carlo techniques is the extensive computational cost and the slow converges properties. Furthermore, due to the stochastic nature of the Monte-Carlo method, the approximations of the distance could vary in different computations.

In this paper, we use unscented transform method proposed by Goldberger, *et al*[5].

We show approximation of $D\big(s_1(x) \parallel s_2(x)\big)$ in (1). Let cluster c_1 contains *d*-dimensional distribution $N_d\left(\boldsymbol{\mu}_m^{(1)}, \boldsymbol{\Sigma}_m^{(1)}\right)$ $(m = 1, \cdots, M)$. Expression formula of c_1 is $\mathrm{s}_1(\boldsymbol{x}) = \boldsymbol{\Sigma}_{m=1}^{M} \boldsymbol{\omega}_m^{(1)} \boldsymbol{p}(\boldsymbol{x}|\boldsymbol{\theta}_m^{(1)})$, where $\boldsymbol{\omega}_m^{(1)}$ is a mixture weight, $p(\boldsymbol{x}|\boldsymbol{\theta}_m^{(1)})$ is *m*-th probability density function of $N_d\left(\boldsymbol{\mu}_m^{(1)}, \boldsymbol{\Sigma}_m^{(1)}\right)$ and $\boldsymbol{\theta}_{\mathrm{m}}^{(1)} = \left(\boldsymbol{\mu}_m^{(1)}, \boldsymbol{\Sigma}_m^{(1)}\right)$. Simmilary, cluster c_2 contains *d*-dimensional distribution $N_d\left(\boldsymbol{\mu}_l^{(2)}, \boldsymbol{\Sigma}_l^{(2)}\right)$ $(l = 1, \cdots, L)$. Expression formula of c_2 is $\mathrm{s}_2(\boldsymbol{x}) = \boldsymbol{\Sigma}_{l=1}^{L} \boldsymbol{\omega}_l^{(2)} \boldsymbol{p}(\boldsymbol{x}|\boldsymbol{\theta}_l^{(2)})$.

Approximation of KL-divergence from s1 to s2 by using unscented transform method is

$$\mathrm{D}(\mathrm{s}_1 || \mathrm{s}_2) \approx \frac{1}{2d} \Sigma_{m=1}^{M} \omega_m \Sigma_{k=1}^{2d} \log \frac{s_1(o_{m,k})}{s_{2(o_{m,k})}} , \qquad (2)$$

where $\mathrm{o}_{\mathrm{m,t}}$ are sigma points. They are chose as follows:

$$\begin{aligned} o_{m,t} &= \mu_m^{(1)} + \left(\sqrt{d\Sigma_m^{(1)}}\right)_t , \\ o_{m,t+d} &= \mu_m^{(1)} + \left(\sqrt{d\Sigma_m^{(1)}}\right)_t \quad , \end{aligned} \qquad (3)$$

such that $\left(\sqrt{\Sigma_m^{(1)}}\right)_t$ is t-th column of the matrix square root of $\Sigma_{\mathrm{m}}^{(1)}$. Then,

$$\mathrm{o}_{\mathrm{m,t}} = \mu_m^{(1)} + \sqrt{d\lambda_{m,t}^{(1)}}\,\boldsymbol{u}_{m,t}^{(1)}$$
$$\mathrm{o}_{\mathrm{m,t}} = \mu_m^{(1)} - \sqrt{d\lambda_{m,t}^{(1)}}\,\boldsymbol{u}_{m,t}^{(1)} \qquad (4)$$

where $t = 1, \ldots, d$, $\mu_m^{(1)}$ is mean vector of m-th normal distribution in s_1, $\lambda_{m,t}^{(1)}$ is t-th eigenvalue of $\Sigma_{\mathrm{m}}^{(1)}$ and $\boldsymbol{u}_{(m,t)}^{(1)}$ is t-th eigenvector. If $p = 1$, the sigma points are simply

$$\mu_m^{(1)} \pm \sigma_m^{(1)}$$

We can calculate approximation of $D(s_2||s_1)$. Substituting these approximations into (1), we obtain the symmetric KL-divergence. We set the divergence as distance between cluster c_1 and c_2.

5.2. Distance Between PDs

In section 5.1, we use symmetric KL-divergence as distance between PDs. It is symmetric KL-divergence between Gaussian mixture distributions. However, it cannot be analytically computed. We can use, instead, Monte-Carlo simulations to approximate the symmetric KL-divergence. The drawback of the Monte-Carlo techniques is the extensive computational cost and the slow converges properties. Furthermore, due to the stochastic nature of the Monte-Carlo method, the approximations of the distance could vary in different computations.

In this paper, we use unscented transform method proposed by Goldberger, *et al*[5].

6. An Application to the VAS Data

In this section, we apply our proposal method to real VAS data from Keio University School of Medicine. This is masked data and is not be tied to any information that would identify a patient.

6.1. Medical Questionnaire in Keio University School of Medicine

Center for Kampo Medicine, Keio University School of Medicine, have a questionnaire to patients to help medical decision. The questionnaire includes one set of questions about their subjective symptoms. There are 244 yes-no questions and 118 visual analogue scale questions,for example, "How do you feel pain with urination?". Patients answer these questions every time when they come to Keio University. Doctors can understand patients' fluctuate in severity.

6.2. Data Description and Result

For our analysis, we deal with four question: "Do you feel cold in your leg?", "Do you feel pain in your leg?", " Do you feel cold in your hand?", "Do you feel pain in your hand?". The data contain 113 patients' first and second VAS value. We transform this data set to PD. And we got some result as figures of dendrogram. The result of our simulation show in figure3. Vertical axis of this dendrogram means distance between PDs.

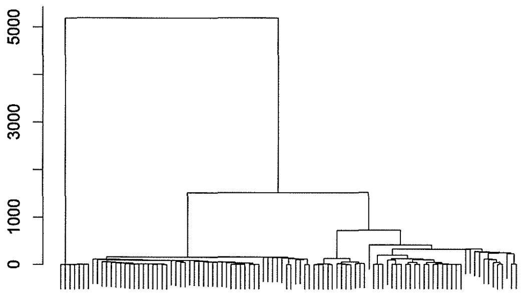

Figure 3 Dendrogram for PDs

7. Concluding Remarks

In this paper, we defined PD that is from transformation of the VAS to Distribution-Valued data. We also proposed hierarchical clustering method for it. Comparing across a group of patients by using the VAS is difficult, but our

method can do it. Through the simulation, we verified our model.
In the future, we will define multidimensional PD and apply our clustering method.

Acknowledgements

This study was partially supported by Health and Labour Sciences Research Grants for Clinical Research from the Ministry of Health, Labour, and Welfare of Japan.

Reference

[1] Bock,H.-H., Diday,E.: Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data. Springer-Verlag, Berlin (2000)
[2] Billard, L., Diday, E.:“ Symbolic Data Analysis ”. Wiley, NewYork (2006)
[3] Dexter F, Chestnut DH. Analysis of statistical tests to compare visual analog scale measurements among groups. Anesthesiology, 82, 896.902.(1995)
[4] Diday, E.: The symbolic approach in clustering and related methods of Data Analysis, In Classification and Related methods Of Data Analysis, H.Bock (ed.), 673.684, Amsterdam: North-Holland (1988)
[5] Goldberger, J., Gordon, S., Greenspan, H.: An efficient image similarity measure based on approximations of KL-divergence between two gaussian mixtures. Proceedings of CVPR, 487.494. (2006)
[6] Gowda, K.C., Diday, E.: Symbolic clustering using a new dissimilarity measure. Pattern Recognition, 24, 6, 567.578. (1991)
[7] Katayama,K., Suzukawa,A., Minami, H., Mizuta, M.: Linearly Restricted Principal Components in k Groups. Electronic proceedings of Knowledge Extraction and Modeling, Villa Orlandi, Island of Capri, Italy (2006)
[8] Katayama,K., Yamaguchi, R., Imoto, S., Tokunaga, H., Imazu, Y., Matuura, K., Watanebe, K., Miyano, S.: Symbolic Hierarchical Clustering for Visual Analogue Scale Data, KES-Springer Smart Innovations, Systems and technologies series, Springer Verlag (2011)
[9] Kullback,S.:“ Information theory and statistics ”, Dover Publications, New York (1968)
[10] Price DD, Bush FM, Long S, Harkins SW. A comparison of pain measurement characteristics of mechanical visual analogue and simple numerical rating scales. Pain, 56, 217.26.(1994)

Symbolic Hierarchical Clustering for Pain Vector

Kotoe Katayama, Rui Yamaguchi, Seiya Imoto, Keiko Matsuura,
Kenji Watanabe, and Satoru Miyano

Abstract. We propose a hierarchical clustering in the framework of Symbolic Data Analysis(SDA). SDA was proposed by Diday at the end of the 1980s and is a new approach for analysing huge and complex data. In SDA, an observation is described by not only numerical values but also "higher-level units"; sets, intervals, distributions, etc. Most SDA works have dealt with only intervals as the descriptions. We already proposed "*pain distribution*" as new type data in SDA. In this paper, we define new "*pain vector*" as new type data in SDA and propose a hierarchical clustering for this new type data.

Keywords: Visual Analogue Scale, Distribution-Valued Data.

1 Introduction

Conventional data analysis usually can handle scalars, vectors and matrices. However, lately, some datasets have grown beyond the framework of conventional data analysis. Most statistical methods do not have sufficient power to analyse these datasets. In this study, we attempted to extract useful information from such datasets.

Symbolic data analysis (SDA) proposed by Diday [3] is an approach for analysing new types of datasets. "Symbolic data" consist of a *concept* that is described by intervals, distributions, etc. as well as by numerical values. The use of SDA enriches data description, and it can handle highly complex datasets. This implies that complex data can be formally handled in the framework of SDA. However, most SDA works have dealt with only intervals as the descriptions and are very few studies

Kotoe Katayama · Rui Yamaguchi · Seiya Imoto · Satoru Miyano
Human Genome Center, Institute of Medical Science, The University of Tokyo,
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
e-mail: `k-kata@ims.u-tokyo.ac.jp`

Keiko Matsuura · Kenji Watanabe
Center for Kampo Medicine, Keio University School of Medicine, 35 Shinano-machi,
Shinjuku-ku, Tokyo 160-8582, Japan

J. Watada et al. (Eds.): Intelligent Decision Technologies, Vol. 2, SIST 16, pp. 117–124.
springerlink.com

based on this simple idea. The case that *concept* is described by intervals is simple, but ignores detailed information in the intervals. We already proposed a hierarchical clustering for the visual analogue scale (VAS) in the framework of Symbolic Data Analysis (SDA)[8].

In this paper, we define *"pain vector"* and propose a hierarchical clustering for this vector. The *"pain vector"* is consist of distributions and categories.

2 Transform the Visual Analogue Scale into *"Patient Distribution"*

We already proposed a hierarchical clustering for the visual analogue scale (VAS) in the framework of Symbolic Data Analysis (SDA)[8]. In the paper, we transformed the Visual Analogue Scale into distribution valued data. The VAS is a method that can be readily understood by most people to measure a characteristic or attitude that cannot be directly measured. VAS is of most value when looking at change within a same people, and is of less value for comparing across a group of people because they have different sense. It could be argued that a VAS is trying to produce interval/ratio data out of subjective values that are at best ordinal. Thus, some caution is required in handling VAS. We described VAS as distribution and handle it as new type data in SDA.

2.1 The Visual Analogue Scale

The visual analogue scale (VAS) has developed to allow the measurement of individual's responses to physical stimuli, such as heat. The VAS is a method that can be readily understood by most people to measure a characteristic or attitude that cannot be directly measured. It was originally used in the field of psychometrics, and nowadays widely used to assess changes in patient health status with treatment.

A VAS consists of a line on a page with clearly defined end points, and normally a clearly identified scale between the two end points. For guidance, the phrase "no pain" and "worst imaginable pain" are placed at the both side of the line, respectively. Minimum value 0 of the VAS means "no pain" and maximum value 100 means "worst imaginable pain".

These scales are of most value when looking at change within patients, and are of less value for comparing across a group of patients because patient have a different sense of pain. It could be argued that a VAS is trying to produce interval/ratio data out of subjective values that are at best ordinal. Thus, some caution is required in handling such data. Many researchers prefer to use a method of analysis that is based on the rank ordering of scores rather than their exact values, to avoid reading too much into the precise VAS score.

2.2 *Transform the Visual Analogue Scale to Distribution-Valued Data*

We transform the VAS to distribution-valued data to compare across a group of patients . VAS varies according to patients, because sense of pain varies a great deal depending on people. Changing VAS score within patients means their sense of pain. If they have big change of VAS score, their expression of sense of pain is rough. On the contrary, if they have small change, their expression is sensitive. We suggest that these sense of pain is described by normal distribution and call it "*pain distribution*(PD)".

Let VAS score of patient's first time be x_1 and second time be x_2. We define the middle point of x_1 and x_2 as mean of PD μ, and $(\mu - x_1)^2 = (\mu - x_2)^2$ as variance. We describe PD as $N(\mu, \sigma^2)$. In case that the number of VAS score is d, PD is d-dimensional normal distribution. In this case, a diagonal matrix is used as a variance-covariance matrix of d-dimensional normal distribution.

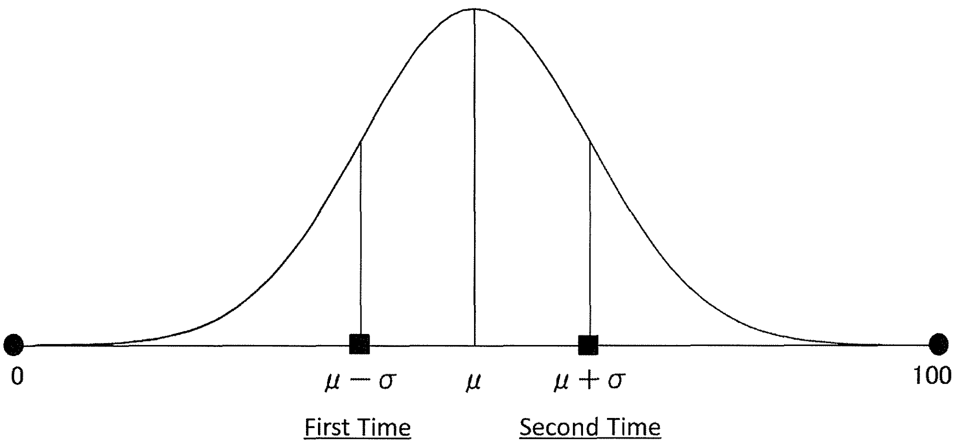

Fig. 1 Transform the Visual Analogue Scale to Distribution-Valued data

3 Medical Questionnaire in Keio University School of Medicine

Center for Kampo Medicine, Keio University School of Medicine, have a questionnaire to patients to help medical decision. The questionnaire includes one set of questions about their subjective symptoms. There are 244 yes-no questions, for example, "Are you constipate?", and 118 visual analogue scale questions, for example, "How do you feel pain with urination?". Patients answer these questions every time when they come to Keio University. Doctors can understand patients' fluctuate in severity.

4 *Pain Vector*

The Medical Questionnaire in Keio University School of Medicine is consist of "*Yes-No*" questions and "*VAS*" questions. To compare "*VAS*" questions among patients, we already proposed “*pain distribution*(PD)”. In the case, we only use ”*VAS*” questions and didn’t use ”*Yes-No*” questions.

Now we propose *Pain Vector* by using both questions. Let ”*Yes-No*” question be YN_a $(a = 1,\dots,A)$. If patient answer ”Yes”, $YN_a = 1$, otherwise 0. Let ”*VAS*” question be PD_b $(b = 1,\dots,B)$. The i-th patient’s *Pain Vector* is

$$\boldsymbol{PV}_i = [YN_{i1},\dots,YN_{iA},PD_{i1},\dots,PD_{iB}]'.$$

5 Hierarchical Clustering for Pain Vector

Cluster analysis groups data objects only on the bases of information found in the data that describes the objects and their relationships. The goal is that the objects within a group should be similar (or related) to one another and different from the objects in other groups.

In this section, we propose a hierarchical clustering for *Pain Vector.*

5.1 The Clustering Algorithm

We extend the idea of a hierarchical clustering in the framework of conventional data analysis. Let n be the number of PV and K be the number of cluster.

<Step1> Begin with K clusters, each containing only a single PV, $K = n$. Calculate distance between PV.

<Step2> Search the minimum distance in K clusters. Let the pair the selected clusters. Combine PVs into a new cluster, it is described by new Vector. Let K be $K-1$. If $K > 1$, go to Step3, otherwise Step4.

<Step3> Calculate the distance between new cluster and other cluster, and go back to Step2.

<Step4> Draw the dendrogram.

5.2 Distance between PVs

In our method, PVs consist of binary and distribution valued data. Let i-th patient’s vector be $\boldsymbol{PV}_i$ and j-th be $\boldsymbol{PV}_j$.

$$\begin{aligned}\boldsymbol{PV}_i &= [\boldsymbol{YN}_i',\boldsymbol{PD}_i']' = [YN_{i1},\dots,YN_{iA},PD_{i1},\dots,PD_{iB}]',\\ \boldsymbol{PV}_j &= [\boldsymbol{YN}_j',\boldsymbol{PD}_j']' = [YN_{j1},\dots,YN_{jA},PD_{j1},\dots,PD_{jB}]'.\end{aligned}$$

The sum of distance between $\boldsymbol{YN}$ and distance between $\boldsymbol{PD}$ is distance between $\boldsymbol{PV}_i$ and $\boldsymbol{PV}_j$.

5.2.1 Distance between YNs

YNs are binary data. We define distance between YN_{ia} and YN_{ja} as

$$||YN_{ia} - YN_{ja}||.$$

Distance between $\boldsymbol{YN}_i$ and $\boldsymbol{YN}_j$ is

$$\sum_{a=1}^{A} ||YN_{ia} - YN_{ja}||.$$

5.2.2 Distance between PDs

We use symmetric KL-divergence as distance between PDs. Kullback-Leibler divergence is the natural way to define a distance measure between probability distributions [9], but not symmetry. We would like to use the symmetric Kullback-Leibler (symmetric KL) divergence as distance. The symmetric KL-divergence between two distributions s_1 and s_2 is

$$\begin{aligned} D(s_1(\boldsymbol{x}), s_2(\boldsymbol{x})) &= D(s_1(\boldsymbol{x})||s_2(\boldsymbol{x})) + D(s_2(\boldsymbol{x})||s_1(\boldsymbol{x})) \\ &= \int_{-\infty}^{\infty} s_1(\boldsymbol{x}) \log \frac{s_1(\boldsymbol{x})}{s_2(\boldsymbol{x})} d\boldsymbol{x} + \int_{-\infty}^{\infty} s_2(\boldsymbol{x}) \log \frac{s_2(\boldsymbol{x})}{s_1(\boldsymbol{x})} d\boldsymbol{x}, \end{aligned} \tag{1}$$

where $D(s_1||s_2)$ is KL divergence from s_1 to s_2 and $D(s_2||s_1)$ is one from s_2 to s_1.

Let PDs be d dimensional $N(\boldsymbol{\mu}_{ib}, \boldsymbol{\Sigma}_{ib})$ and $N(\boldsymbol{\mu}_{jb}, \boldsymbol{\Sigma}_{jb})$. Symmetric KL-divergence is

$$\begin{aligned} &D(p(\boldsymbol{x}|\boldsymbol{\mu}_{ib}, \boldsymbol{\Sigma}_{ib}), p(\boldsymbol{x}|\boldsymbol{\mu}_{jb}, \boldsymbol{\Sigma}_{jb})) \\ &\qquad = tr(\boldsymbol{\Sigma}_{ib}\boldsymbol{\Sigma}_{jb}^{-1}) + tr(\boldsymbol{\Sigma}_{jb}\boldsymbol{\Sigma}_{ib}^{-1}) \\ &\qquad\quad + tr((\boldsymbol{\Sigma}_{ib}^{-1} + \boldsymbol{\Sigma}_{jb}^{-1})(\boldsymbol{\mu}_{ib} - \boldsymbol{\mu}_{jb})(\boldsymbol{\mu}_{ib} - \boldsymbol{\mu}_{jb})^T) - 2d. \end{aligned} \tag{2}$$

Let PDs be $d = 1$,

$$\begin{aligned} &D(p(x|\mu_{ib}, \sigma_{ib}), p(x|\mu_{jb}, \sigma_{jb})) \\ &= \frac{1}{2}\left\{\log \frac{\sigma_{jb}^2}{\sigma_{ib}^2} + \frac{\sigma_{ib}^2 + (\mu_{ib} - \mu_{jb})^2}{\sigma_{jb}^2}\right\} + \frac{1}{2}\left\{\log \frac{\sigma_{ib}^2}{\sigma_{jb}^2} + \frac{\sigma_{jb}^2 + (\mu_{jb} - \mu_{ib})^2}{\sigma_{ib}^2}\right\} -(\mathrm{B.}) \end{aligned}$$

Distance between $\boldsymbol{PD}_i$ and $\boldsymbol{PD}_j$ is

$$\sum_{b=1}^{B} D(p(x|\mu_{ib}, \sigma_{ib}), p(x|\mu_{jb}, \sigma_{jb})).$$

5.3 *New PV in Clustering Algorithm Step 2 and Their Distance*

In Clustering Algorithm Step 2, we combine $\boldsymbol{PV}_i$ and $\boldsymbol{PV}_j$ into a new cluster and it is described by new vector. This new vector, $\boldsymbol{NPV}$, is described by using distance between YN_{ia} and YN_{ja} and Gaussian mixture distributions of PD_{ib} and PD_{jb}.

$$\boldsymbol{NPV} = [YN_{i1} - YN_{j1}, \ldots, YN_{iA} - YN_{jA}, mgd(PD_{i1}, PD_{j1}), \ldots, mgd(PD_{iB}, PD_{jB})]',$$

where $mgd(PD_{i1}, PD_{j1})$ means mixture distribution of PD_{i1} and PD_{j1}, and mixture weight equal 0.5.

After Section 5.1 Step2, we need symmetric KL-divergence between Gaussian mixture distributions. However, it cannot be analytically computed. We can use, instead, Monte-Carlo simulations to approximate the symmetric KL-divergence. The drawback of the Monte-Carlo techniques is the extensive computational cost and the slow converges properties. Furthermore, due to the stochastic nature of the Monte-Carlo method, the approximations of the distance could vary in different computations.

In this paper, we use unscented transform method proposed by Goldberger, *et al*[5].

We show approximation of $D(s_1||s_2)$ in (1). Let cluster c_1 contains d-dimensional distribution $N_d(\boldsymbol{\mu}_m^{(1)}, \boldsymbol{\Sigma}_m^{(1)})$, $(m = 1, \ldots M)$. Expression formula of c_1 is $s_1(\boldsymbol{x}) = \sum_{m=1}^{M} \omega_m^{(1)} p(\boldsymbol{x}|\boldsymbol{\theta}_m^{(1)})$, where $\omega_m^{(1)}$ is a mixture weight, $p(\boldsymbol{x}|\boldsymbol{\theta}_m^{(1)})$ is m-th probability density function of $N_d(\boldsymbol{\mu}_m^{(1)}, \boldsymbol{\Sigma}_m^{(1)})$ and $\boldsymbol{\theta}_m^{(1)} = (\boldsymbol{\mu}_m^{(1)}, \boldsymbol{\Sigma}_m^{(1)})$. Simmilary, cluster c_2 contains d-dimensional distribution $N_d(\mu_l^{(2)}, \Sigma_l^{(2)})(l = 1, \ldots L)$. Expression formula of c_2 is $s_2 = \sum_{l=1}^{L} \omega_n^{(2)} p(\boldsymbol{x}|\boldsymbol{\theta}_l^{(2)})$.

Approximation of KL-divergence from s_1 to s_2 by using unscented transform method is

$$D(s_1||s_2) \approx \frac{1}{2d} \sum_{m=1}^{M} \omega_m \sum_{k=1}^{2d} \log \frac{s_1(\boldsymbol{o}_{m,k})}{s_2(\boldsymbol{o}_{m,k})}, \tag{4}$$

where $\boldsymbol{o}_{m,t}$ are sigma points. They are chose as follows:

$$\begin{aligned} \boldsymbol{o}_{m,t} &= \boldsymbol{\mu}_m^{(1)} + \left(\sqrt{d\boldsymbol{\Sigma}_m^{(1)}}\right)_t, \\ \boldsymbol{o}_{m,t+d} &= \boldsymbol{\mu}_m^{(1)} - \left(\sqrt{d\boldsymbol{\Sigma}_m^{(1)}}\right)_t, \end{aligned} \tag{5}$$

such that $\left(\sqrt{\boldsymbol{\Sigma}_m^{(1)}}\right)_t$ is t-th column of the matrix square root of $\Sigma_m^{(1)}$. Then,

$$\begin{aligned} \boldsymbol{o}_{m,t} &= \boldsymbol{\mu}_m^{(1)} + \sqrt{d\lambda_{m,t}^{(1)}}\boldsymbol{u}_{m,t}^{(1)} \\ \boldsymbol{o}_{m,t+d} &= \boldsymbol{\mu}_m^{(1)} - \sqrt{d\lambda_{m,t}^{(1)}}\boldsymbol{u}_{m,t}^{(1)}, \end{aligned} \tag{6}$$

where $t = 1, \ldots, d$, $\boldsymbol{\mu}_m^{(1)}$ is mean vector of m-th normal distribution in s_1, $\lambda_{m,t}^{(1)}$ is t-th eigenvalue of $\boldsymbol{\Sigma}_m^{(1)}$ and $\boldsymbol{u}_{m,t}^{(1)}$ is t-th eigenvector. If $d = 1$, the sigma points are simply

$$\mu_m^{(1)} \pm \sigma_m^{(1)}.$$

We can calculate approximation of $D(s_2||s_1)$. Substituting these approximations into (1), we obtain the symmetric KL-divergence. We set the divergence as distance between PD in cluster c_1 and PD in c_2.

6 An Application

In this section, we apply our proposal method to real data from Keio University School of Medicine. This is masked data and is not be tied to any information that would identify a patient. For our analysis, we use the 2316 patients' result of medical questionnaire. There are 244 yes-no questions, and 118 visual analogue scale questions.

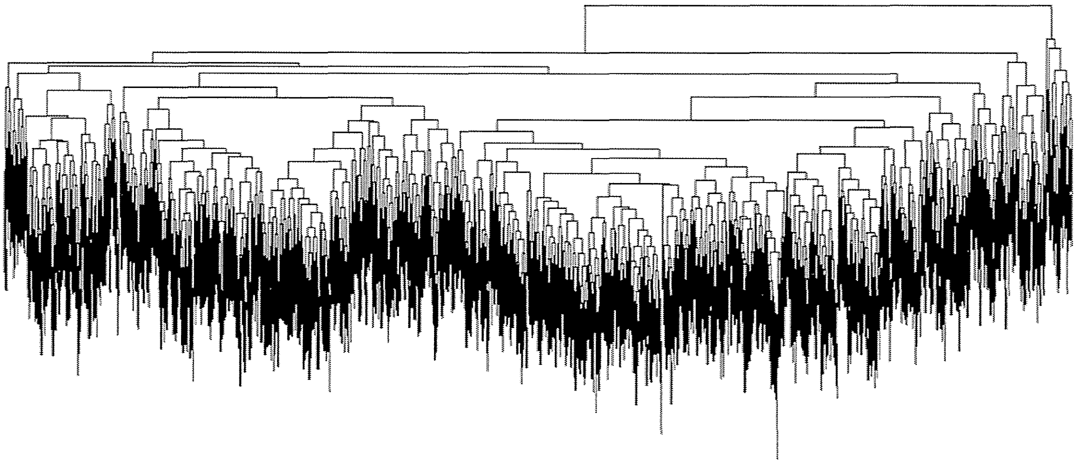

Fig. 2 Dendrogram for 2316 patients

7 Concluding Remarks

We already proposed a hierarchical clustering for the visual analogue scale (VAS) in the framework of Symbolic Data Analysis (SDA). To compare "*VAS*" questions among patients, we already proposed "*pain distribution*(PD)". In the case, we only used "*VAS*" questions and didn't use "*Yes-No*" questions. In this paper, we use both questions and define new "*pain vector*" as new type data in SDA. The "*pain vector* consist of binary and distribution valued data. We also propose a hierarchical clustering for this new type data. Through the simulation, we verified our model.

References

1. Billard, L., Diday, E.: Symbolic Data Analysis. Wiley, NewYork (2006)
2. Bock, H.-H., Diday, E.: Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data. Springer, Berlin (2000)

3. Diday, E.: The symbolic approach in clustering and related methods of Data Analysis. In: Bock, H. (ed.) Proc. IFCS Classification and Related Methods of Data Analysis, Aachen, Germany. North-Holland (1988)
4. Diday, E.: The symbolic approach in clustering and related methods of Data Analysis. In: Bock, H. (ed.) Classification and Related methods of Data Analysis, pp. 673–684. North-Holland, Amsterdam (1988)
5. Goldberger, J., Gordon, S., Greenspan, H.: An efficient image similarity measure based on approximations of KL-divergence between two gaussian mixtures. In: Proceedings of CVPR, pp. 487–494 (2006)
6. Gowda, K.C., Diday, E.: Symbolic clustering using a new dissimilarity measure. Pattern Recognition 24(6), 567–578 (1991)
7. Katayama, K., Suzukawa, A., Minami, H., Mizuta, M.: Linearly Restricted Principal Components in *k* Groups. In: Electronic proceedings of Knowledge Extraction and Modeling, Villa Orlandi, Island of Capri, Italy (2006)
8. Katayama, K., Yamaguchi, R., Imoto, S., Matsuura, K., Watanabe, K., Miyano, S.: Clustering for Visual Analogue Scale Data in Symbolic Data Analysis. Procedia Computer Science 6, 370–374 (2011)
9. Kullback, S.: Information theory and statisticsh. Dover Publications, New York (1968)

Connection between Traditional Medicine and disease

Kotoe Katayama
Institute of Medical Science,
University of Tokyo
4-6-1 Shirokanedai, Minato-ku Tokyo,
Japan 108-8639
k-kata@ims.u-tokyo.ac.jp

Rui Yamaguchi
Institute of Medical Science,
University of Tokyo
4-6-1 Shirokanedai, Minato-ku Tokyo,
Japan 108-8639

Seiya Imoto
Institute of Medical Science,
University of Tokyo
4-6-1 Shirokanedai, Minato-ku Tokyo,
Japan 108-8639

Keiko Matsuura
Center of Kampo Medicine,
Keio University School of Medicine
35 Shinano-machi, Shinjyuku-ku Tokyo,
Japan 160-8639

Kenji Watanabe
Center of Kampo Medicine,
Keio University School of Medicine
35 Shinano-machi, Shinjyuku-ku Tokyo,
Japan 160-8639

Seiya Imoto
Institute of Medical Science,
University of Tokyo
4-6-1 Shirokanedai, Minato-ku Tokyo,
Japan 108-8639

ABSTRACT

In Japanese traditional medicine, “Monshin” plays an important role. “Monshin” is a questionnaire that asked the patient's lifestyle and subjective symptoms. Specialists decide traditional herbal medicine by using of “Monshin”. In this research, we connect “Monshin” to disease through building the Network.

Keywords

traditional Medicine, regression model, elastic net, network.

1. INTRODUCTION

Interest in traditional medicine has increased globally in recent years. The World Health Organization (WHO) recommends the use of traditional medicine for care and treatment of people with health conditions. Further, the WHO suggests the integration of traditional medicine into the next edition of the International Statistical Classification of Diseases and Related Health Problems (ICD-11).

Kampo medicine -the Japanese adaptation of traditional medicine-originated in ancient China and arrived in Japan via the Korean peninsula in the 6th century, after which it developed uniquely. Kampo medicine was approved in 1967 under the National Health Insurance policy. Of the 50 years of National Health Insurance for whole nations, 44 years dealt with Japanese traditional medicine.

A remarkable feature of the Japanese medical licence is that a doctor with that license is allowed to combine modern and Kampo medicine for a medical treatment and prescribe both kinds of drugs; while in most of other countries, e.g., China and Korea, they have different licenses for modern and traditional medicine, thus, it is hard to utilize them in a combinatorial way.

2. “Monshin” NETWORK

In Japanese traditional medicine, “Monshin” plays an important role. “Monshin” is a questionnaire that asked the patient's lifestyle and subjective symptoms. Specialists decide traditional herbal medicine by using of “Monshin”.

However, in order to determine herbal medicine or traditional cure, technical knowledge and experience are required. In Keio University School of Medicine, we analyze “Monshin” data to establish an indicator for non Kampo specialist without technical knowledge to perform suitable traditional medicine. In this research, we connect “Monshin” to disease through building the “Monshin” Network. To build it, we identified items of “Monshin” relevant to a disease by using of logistic regression model and elastic net.

3. RESULT

We focus atopy as a target disease and build the “Monshin” Network and show it in figure 1 and 2.

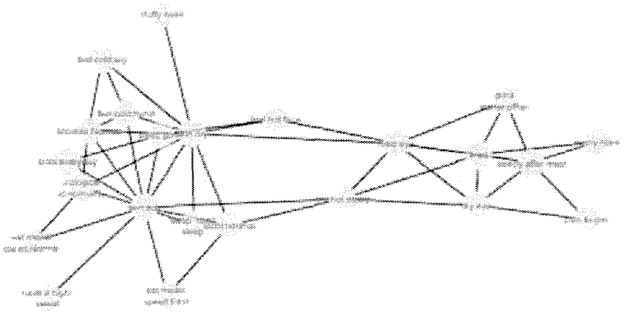

Figure 1. Monshin Network for Atopy (positive)

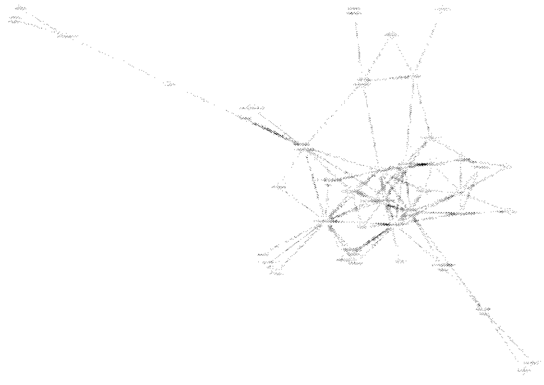

Figure 2. Monshin Network for Atopy (negative)

4. ACKNOWLEDGMENTS

This study was partially supported by Health and Labour Sciences Research Grants for Clinical Research and Research on Statistics and Information from the Ministry of Health, Labour, and Welfare of Japan.

Analysis of questionnaire for Traditional Medical and develop decision support system

Kotoe Katayama
Institute of Medical Science
University of Tokyo
4-6-1 Shirokanedai, Minato-ku Tokyo, Japan 108-8639
k-kata@ims.u-tokyo.ac.jp

Rui Ymaguchi
Institute of Medical Science
University of Tokyo
4-6-1 Shirokanedai, Minato-ku Tokyo, Japan 108-8639

Seiya Imoto
Institute of Medical Science
University of Tokyo
4-6-1 Shirokanedai, Minato-ku Tokyo, Japan 108-8639

Keiko Matsuura
Center of Kampo Medicine,
Keio University School of Medicine
35 Shinano-machi, Shinjyuku-ku Tokyo, Japan 160-8582

Kenji Watanabe
Center of Kampo Medicine,
Keio University School of Medicine
35 Shinano-machi, Shinjyuku-ku Tokyo, Japan 160-8582

Satoru Miyano
Institute of Medical Science
University of Tokyo
4-6-1 Shirokanedai, Minato-ku Tokyo, Japan 108-8639

Abstract*—*In Japanese traditional medicine, "Monshin" plays an important role. "Monshin" is a questionnaire that asked the patient's lifestyle and subjective symptoms. Specialists decide traditional herbal medicine by using of "Monshin". In this research, we analyze "Monshin" and predict "Sho" which is the name of a disease.

***Keywords-** Traditional medicine; random forest; machine learning;*

I. INTRODUCTION

Interest in traditional medicine has increased globally in recent years. The World Health Organization (WHO) suggests the integration of traditional medicine into the next edition of the International Statistical Classification of Diseases and Related Health Problems (ICD-11).

Kampo medicine -the Japanese adaptation of traditional medicine- originated in ancient China and arrived in Japan via the Korean peninsula in the 6th century, after which it developed uniquely. A remarkable feature of the Japanese medical license is that a doctor with that license is allowed to combine modern and Kampo medicine for a medical treatment and prescribe both kinds of drugs; while in most of other countries, e.g., China and Korea, they have different licenses for modern and traditional medicine, thus, it is hard to utilize them in a combinatorial way.

In Japanese traditional medicine, "Monshin" plays an important role. "Monshin" is a questionnaire that asked the patient's lifestyle and subjective symptoms. Specialists decide traditional herbal medicine by using of "Monshin". However, in order to determine herbal medicine or traditional cure, technical knowledge and experience are required. In this paper, we analyze "Monshin" data to establish an indicator for non-Kampo specialist without technical knowledge to perform suitable traditional medicine.

II. DIAGNOSIS BY KAMPO

The diagnosis by Kampo is called "Sho" and determined by completely different view from Western medicine. Kampo uses a unique cognitive paradigm, making use of tools such as "Deficiency and Excess", "Cold and Heat", Qi, Blood, Fluid. The diagnosis by Kampo is composed one item from each "Deficiency and Excess" category and "Cold and Heat" category, and two items from "Qi, Blood, Fluid" category. Specialists will gather all available information to best determine patients' "Sho". There are some types of Kampo examinations: "Monshin", tongue diagnosis, Palpation, and so on. Kampo specialists consider all the various factors together. However it causes difficulties for non-Kampo specialist. It is hard for them to connect result of "Monshin" and other diagnosis.

III. DATA OF KEIO UNIVERSITY

Since 2006, Center of Kampo medicine, Keio university school of medicine has been collected data about patients' "Monshin", "Sho", western disease name (ICD-10 code), and prescribed herbal medicine. Patients enter "Monshin" information via touch panel operation. "Monshin" has 362 items, ranges in content from physical sign to food preference and is important for Kampo diagnosis. There are

978-1-4673-2747-3/12/$31.00 ©2012 IEEE

two type questions, yes-no question and VAS question. From April, 2006 till December, 2011, we collected 16805 records, and the number of first visit patients was 2830.

IV. Predict of "Sho" and Results

We predict "Sho" by using 2830 first visit patients' "Monshin" data. In this paper, we focus on Deficiency and Excess category as a target and adopt random forests.

Random forests was proposed by Breiman and is an algorithm for classification that uses an ensemble of classification trees. Our case is supervised learning. We set training and test data that has labels consistent with that type of classification.

TABLE I. "Sho" of Deficiency and Excess category: 2830 first visit patients

Deficiency and Excess	patients
Deficiency pattern	437
Slightly deficiency pattern	395
Between deficiency and excess	1500
Slightly excess pattern	268
Excess pattern	230
Total	2830

Our target data is Table I. We selected randomly 200 patients as a training data (each 100 patients from deficiency pattern and excess pattern). And others are test data.

TABLE II. Result of test data

		Deficiency pattern	Excess pattern	Discriminant ratio
Predict	Deficiency pattern	231	48	67.0%
	Excess pattern	106	82	
	Total	337	130	

The discriminant ratio of training data was perfect but of test data is 67.0%. It was far from practical use (Table II).

V. Predict of "Sho" with Body Mass Index and Its Result

To cover the shortcomings of our questionnaire, we added Body Mass Index (BIM) data to "Monshin" data. BMI is a simple index of weight-for-height that is commonly used to classify underweight, overweight and obesity in adults. It is defined as the weight in kilograms divided by the square of the height in meters.

Center of Kampo medicine, Keio university school of medicine has patients' BMI data on 2011 year (Table VIII). We selected randomly 40 patients as a training data (each 20 patients from deficiency pattern and excess pattern). And others are test data.

TABLE III. "Sho" of Deficiency and Excess category: year 2011

Deficiency and Excess	patients
Deficiency pattern	75
Slightly deficiency pattern	28
Between deficiency and excess	223
Slightly excess pattern	39
Excess pattern	37
Total	402

The discriminant ratio of training data was perfect and of test data is 91.2% (Table IV).

TABLE IV. Result of test data

		Deficiency pattern	Excess pattern	Discriminant ratio
Predict	Deficiency pattern	51	2	91.2%
	Excess pattern	4	15	
	Total	55	17	

VI. Concluding Remarks

In Japanese traditional medicine, Kampo, "Monshin" plays an important role. The diagnosis by Kampo is called "Sho" and determined by completely different view from Western medicine. And this is reason why non-Kampo specialists without technical knowledge are feel difficulties to use traditional medicine. Since 2006, Center of Kampo medicine, Keio university school of medicine has been collected 2830 first visit patients' data. We predict "Sho" by using Random forests which is powerful algorithm for classification. First, we use all the 2830 first visit patients' data. The discriminant ratio of training data was perfect but of test data is 67.0%. Originally, deficiency and excess category means that patient is strongly built or poor built and our "Monshin" didn't include such indicator. If we use only BMI for classification, it is not working well. So we use both "Monshin" and BMI, and the discriminant ratio of test data is 91.2%. To get good classification, we have to know what is feature of the target and check the data.

In this research, prediction of deficiency and excess category is enough for practical use, but other categories are remained and are our future targets.

Acknowledgment

This study was partially supported by Health and Labour Sciences Research Grants for Clinical Research and Research on Statistics and Information from the Ministry of Health, Labour, and Welfare of Japan.

References

[1] Stafford L. HerbalEGram: Volume 8, Number 1, January 2011 WHO Developing New Traditional Medicine Classification http://cms.herbalgram.org/heg/volume8/01January/WHOClassifiesTM.html?t=1294841964 (accessed April 10, 2012).

[2] Watanabe K, Matsuura K, Gao P, et al. Traditional Japanese Kampo medicine: clinical research between modernity and traditional medicine-the state of research and methodological suggestions for the future, Evid Based Complement Alternat Med 2010; published on line June 16. DOI:10.1093/ecam/neq067.

[3] L. Breiman. Bagging predictors. Machine Learning, 24 (2):123–140, 1996.

Revealing Modern History of Japanese Philosophy Using Natural Language Processing and Visualization

Hideki Mima, Katsuya Masuda, Susumu Ota, and Shunya Yoshimi
Center for Knowledge Structuring, University of Tokyo

Type of presentation: Poster
Keywords: Natural language processing, visualization, Japanese philosophy, thoughts, knowledge structuring
Contact email address: mima@t-adm.t.u-tokyo.ac.jp
Postal address: 7-3-1 Hongou Bunkyou-ku Tokyo 113-8656, Japan

Abstract

The purpose of this study was to reveal the modern history of Japanese philosophy using natural language processing (NLP) and visualization. Knowledge[1] has been increasing at an exponential rate with advances in science and technology in recent years resulting in massive amounts of knowledge that have been extremely difficult to process manually. Thus, it is important to utilize information technologies (IT) to support new discoveries of knowledge from large numbers of resources, such as literature. To implement the study, we have developed:

1) A corpus representing a modern history of Japanese philosophy,

2) A computational model for extracting ontology[2] from the corpus, and

3) An interactive user interface (UI) to support new discoveries of knowledge.

We chose “Shisou” (thoughts) by the Japanese publisher Iwanami Shoten for the target corpus, which is one of the most representative journals of philosophy in Japan that has an almost 90 year history from 1921 to the present-day. It is comprised of about 8,600 papers and more than 160,000 pages of textual data. The first step in this study was to develop a technology to digitize such large amounts of textual data from physical books (semi-) automatically. Because the target was too huge to digitize manually (i.e.,

[1] Although the definition of knowledge is domain-specific, our definition of knowledge here is the particles represented by ontology, which is the (hierarchical) collection and classification of (technical) terms used to recognize their semantic relevance.

[2] Although the definition of ontology is also domain-specific, our definition of ontology h ere is, as previously mentioned, the (hierarchical) collection and classification of (technical) terms used to recognize their semantic relevance.

by typing), a rapid, accurate and low-cost approach was required. Thus, we developed an Optical Character Reader (OCR) based (semi-) automatic book-digitizing system, in which we integrated three processes:

i) Book scanning
ii) OCR
iii) Automatic document style recognition

The input for the system were physical books and the output was a full-text corpus with meta-data, i.e. titles, authors, page numbers, and dates.

We propose a knowledge structuring (KS) system[1] to integrate NLP and the visualization-based interactive UI for the model of ontology extraction and UI. The system architecture is modular, and it integrates five components (Fig. 1): a) information (ontology) extraction, b) a corpus database, c) information retrieval, d) similarity calculations, and e) visualization.

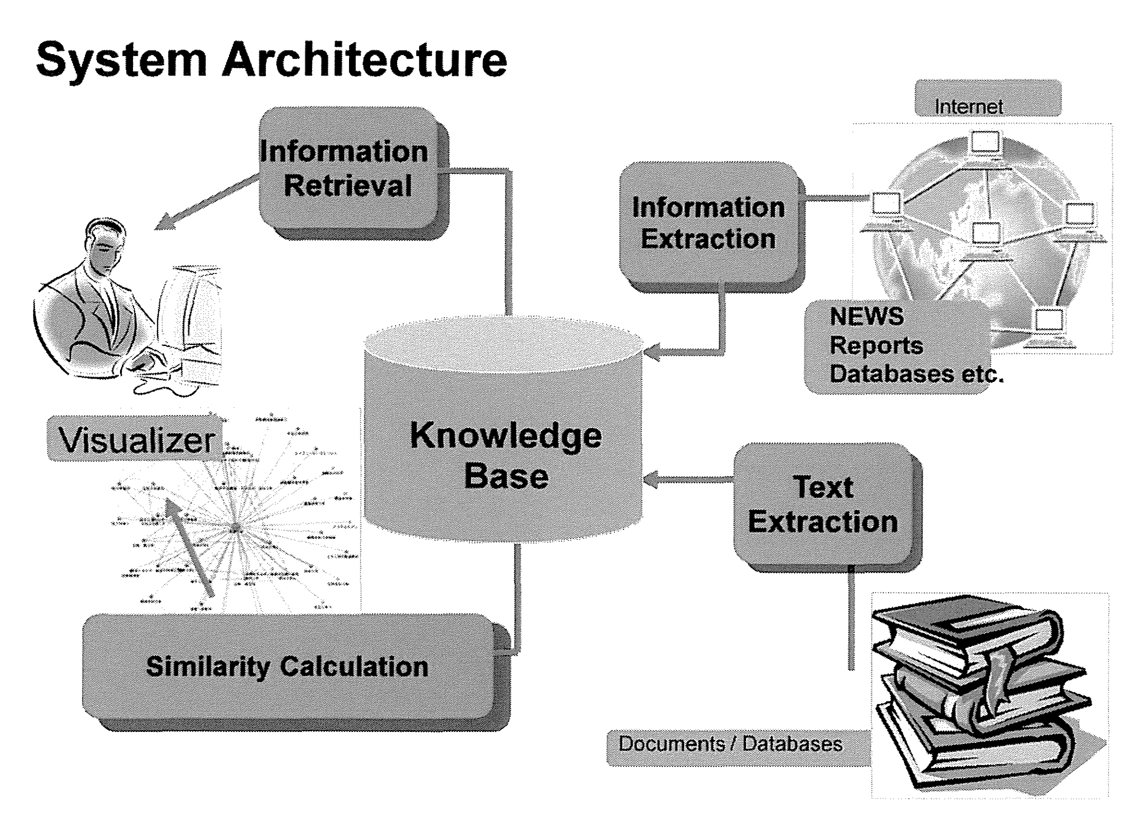

Fig.1 KS System Architecture

The main objective of the system was to facilitate knowledge acquisition from documents and generate ideas through terminology-based real-time calculations of document similarities and their visualization with an interactive UI. Fig. 2 outlines the visualization of knowledge structures for *shisou* papers relevant to the keyword "shisou (thoughts)" in the 1930s. The system constructs a graph to structure knowledge in which the nodes (dots) reflect relevant papers with the keyword, and the links between the nodes reflect semantic similarities that are calculated based on terminological information in the papers. Additionally, the locations of all nodes are calculated and optimized when the graph is drawn. The distance between each node depends on how close they are in meaning. Cluster recognition is also carried out based on the detection of groups of papers in which every combination of papers that are included is strongly linked (i.e., their similarity exceeds a threshold). As seen in Fig. 2, several clusters are automatically recognized and category names such as “Marxism”, “socialism” and “right-wing thoughts” are also automatically assigned to clusters to facilitate an overview of thoughts discussed in these papers.

We have currently finished digitizing and creating a “Shisou” textual database of the 20 years from 1940 to 1959 and installed it in the KS system. Several experiments on text digitization were conducted to evaluate the OCR and style recognition process to improve accuracy. We obtained more than 98% accuracy in OCR, about 90% accuracy in style recognition according to the latest evaluation.

We expect to discover new knowledge on the historical flow of Japanese thinking during one of its most important eras from before World War II to the present-day by digitizing and analyzing huge amounts of historical textual data with the system.

References

[1] Mima, H. and Ananiadou, S. "An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese." *International Journal on Terminology*, 6 (2), pp. 175–194, 2000.

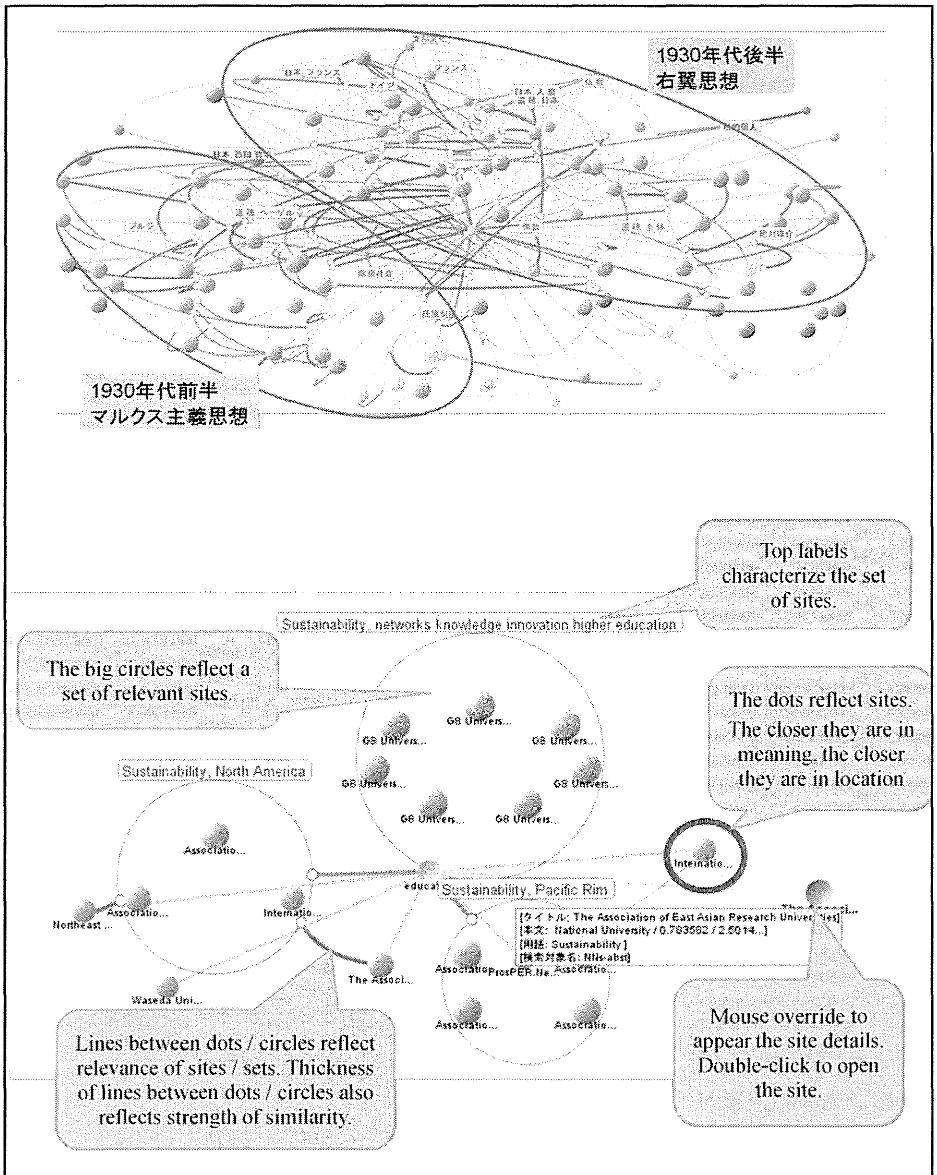
1930年代後半
右翼思想
1930年代前半
マルクス主義思想
日本, フランス
ドイツ
フランス
Top labels characterize the set of sites.
Sustainability, networks knowledge innovation higher education
The big circles reflect a set of relevant sites.
G8 Univers...
Sustainability, North America
Associatio...
Northeast ...
Internatio...
Waseda Uni...
The Associ...
educa
Sustainability, Pacific Rim
The dots reflect sites. The closer they are in meaning, the closer they are in location
[タイトル: The Association of East Asian Research Univer ties]
[本文: National University / 0.783582 / 2.5014...]
[用語: Sustainability]
[検索対象名: NIIs-abst]
ProsPER.Ne...
Mouse override to appear the site details. Double-click to open the site.
Lines between dots / circles reflect relevance of sites / sets. Thickness of lines between dots / circles also reflects strength of similarity.

Fig.2 Visualization of knowledge structures

生命科学における知の構造化

東京大学大学院 工学系研究科（兼任）東京大学 知の構造化センター　美馬 秀樹

Key words　知の構造化 ／ 自然言語処理 ／ 可視化 ／ オントロジー ／ 漢方医学

増え続ける“知識”

1800万件／3万件……、これらはそれぞれ、MEDLINE(医学・生命科学分野の文献データベース)に登録されている文献数、および毎月のおおよその増加数(2010年時)である。ICT(情報通信技術)の発展、科学の拡大、専門分野の深化を背景に、生命科学分野のみならず、あらゆる分野において知識の量が爆発的に増加しており、非専門家はもとより専門家にとっても知識の全体像の把握が非常に困難な状況となっている。さらには、環境やエネルギーのような地球規模での複雑で多様な問題が顕現化し、学際的、分野横断的に知識の活用を促す仕組みの構築がより重要性を増している。

東京大学 知の構造化センターでは、このように自律分散的に創造される膨大な知識を構造化し、新しい知的価値、経済的価値、社会的価値、文化的価値に結びつける「知の構造化」の研究開発を進めている。「知の構造化」により、時勢や学問分野間、また、人や組織間を越えた知識の「インターフェーシング」を行い、知の要素と要素の関係からその全体像を明らかにすることで、多様な知を関連づけ、新しい価値を創出することを目指している。

例えば、医療において、近年、医学と工学の連携により発明された技術として、「3次元血管造影診断技術」がある。血管造影は心筋梗塞の重症度診断等、さまざまな診断で利用されているが、従来は腕や大腿部の動脈から細い管(カテーテル)を入れて造影剤を流し込み、映画撮影するものであった。これは、治療法の選択等にも欠くことのできない有用な検査であるが、患者の時間的、体力的な負担が大きく、簡単にくり返して行えるものではなかった。

これに対し、「3次元血管造影診断」は、ITによる高速センシングと3次元CG(Computer Graphics)を利用した可視化技術により、短時間に検査を行うことができ、患者の負担軽減の観点からもその価値は計り知れない。この発明は、図１に示すように、「医療」と「情報工学」に係る知の構造化、さらには「造影」と「可視化」という知の合成なしには、なし得なかったものである。

本稿では、このような、知識を効

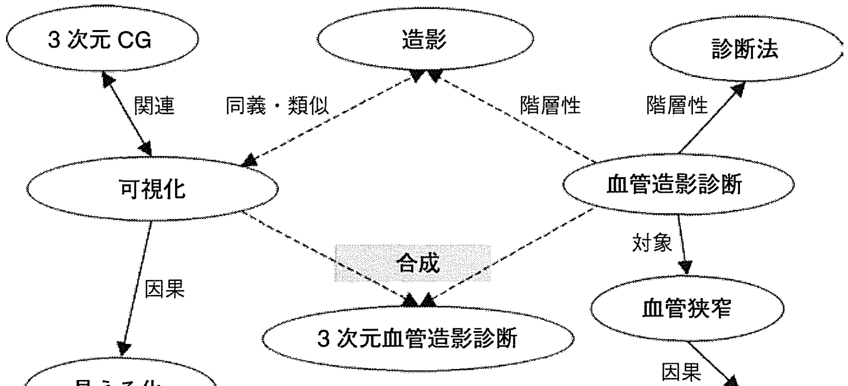

図１　生命科学における知の構造化の例

「情報」「知識」、および「知」という言葉は、ドメインや文脈、状況等によりさまざまな意味を持つ。例えば、Webを対象とした情報抽出は、HTML文書から特定の部分のテキストを抽出することを指す場合が多いが、自然言語処理では、さらにテキストから固有名詞等の特定の情報を抽出することを指す。よって、本稿においても、それらを厳密に定義しないが、知の構造化の対象をテキストとした際の「知」および「知識」の対象としては、(専門)用語等の属性により特徴付けられたパーティクル（文、段落、節、文書等の単位、またはそれらと関連付けられたコンテンツ）を示すものとする。