# USING TEXT AND VISUAL MINING TO ANALYZE CLINICAL DIAGNOSIS RECORDS

*Chien-Hsing Chen*[*]*, Chung-Chian Hsu*[†]

**Abstract:** Hospitals must index each case of inpatient medical care with codes from the International Classification of Diseases, 9th Revision (ICD-9), under regulations from the Bureau of National Health Insurance. This paper aims to investigate the analysis of free-textual clinical medical diagnosis documents with ICD-9 codes using state-of-the-art techniques from text and visual mining fields. In this paper, ViSOM and SOM approaches inspire several analyses of clinical diagnosis records with ICD-9 codes. ViSOM and SOM are also used to obtain interesting patterns that have not been discovered with traditional, nonvisual approaches. Furthermore, we addressed three principles that can be used to help clinical doctors analyze diagnosis records effectively using the ViSOM and SOM approaches. The experiments were conducted using real diagnosis records and show that ViSOM and SOM are helpful for organizational decision-making activities.

Key words: *Clinical diagnosis record, ICD-9 code, keyword extraction, ViSOM, SOM*

*Received: July 16, 2009*
*Revised and accepted: September 26, 2012*

## 1. Introduction

Under regulations from the Bureau of National Health Insurance, hospitals must index each case of inpatient medical care with codes from the International Classification of Diseases, 9th Revision (ICD-9), when applying for reimbursement of medical fees. The ICD-9 code determines the reimbursement amount, and incorrect codes may result in fines. Thus, it is important to conduct the task of code assignment as accurately as possible [Larkey and Croft 1996]. The current coding process used by medical organizations is for physicians to manually assign the initial code

[*]Chien-Hsing Chen
Department of Information Management, Ling Tung University, E-mail: ktfive@gmail.com
[†]Chung-Chian Hsu
Department of Information Management, National Yunlin University of Science and Technology, E-mail: hsucc@yuntech.edu.tw

when writing either the discharge summary or the clinical medical diagnosis. Before submission for reimbursement, the coding is manually checked again by other specialists to ensure correctness and avoid fines. Because the manual process is labor-intensive and time-consuming, an expert medical system that supports the analysis of clinical medical diagnosis records has been in great demand. The studies in the existing literature attempt to address the use of automatic code assignment [Franz, Zaiss et al. 2000], rule extraction [Soualmia and Darmoni 2005; Cerrito and Cerrito 2006], automatic annotation [Boeckmann, Bairoch et al. 2003] and document retrieval [Wilbur and Coffee 1994; Schuler 1996]. However, visual mining and analysis of free-textual medical documents with ICD-9 codes have received relatively little attention in past studies. This paper investigates the analysis of free-textual clinical diagnosis documents with ICD-9 codes using state-of-the-art techniques from text and visual mining fields.

The self-organizing map (SOM), proposed by Kohonen [Kohonen 1989], is an unsupervised neural network that transforms high-dimensional data onto a low-dimensional (usually 2D or 3D) space while preserving the topological order of the original data. In the SOM approach, similar objects in the original data space are projected to the same neuron or nearby neurons in the visualized latent space. Due to this visualization property, the SOM approach has been applied to many applications [Kohonen, Kaski et al. 2000; Hsu 2006; Liu, Wang et al. 2008], such as visual multivariate data cluster analysis [Yin 2002] and visual data clustering.

In this paper, we expect to obtain interesting patterns that have not been discovered with traditional, nonvisual approaches. Moreover, we address three principles that can be used to help users (e.g., clinical doctors) analyze clinical diagnosis records effectively. The approaches are based on SOM [Kohonen 1989] and ViSOM [Yin 2002; Yin 2002]. First, the visualization resulting from the SOM-based approach provides clinical doctors with an understanding of the context of clinical diagnosis records. Second, SOM-based approaches visually show the relationships between neurons (nodes) on the map because all high-dimensional data have already been projected onto the low-dimensional space. For example, clinical doctors can easily appreciate the fact that diagnosis records projected near other nodes are similar to those nodes and less similar to nodes that are farther away on the map. Finally, exploring free-textual documents with map neurons can support searches for relevant documents. The search can start at a satisfactory map neuron and then expand to neighboring map neurons [Kohonen, Kaski et al. 2000]. These three applications rely on projecting real diagnosis records onto two-dimensional maps using ViSOM and SOM. We analyze the visual use scenario created by the projection through a comparison of coincident qualitative and quantitative measurements for these three applications.

The rest of the paper is organized as follows. After Section 1 (Introduction), Section 2 illustrates the related work. The analysis of clinical diagnosis records with ICD-9 codes is introduced in Section 3. The visual cluster analysis methods for diagnosis records are introduced in Section 4. Section 5 describes the experiments, and Section 6 contains discussions and conclusions.

## 2. Related Work

Over the last decade, biomedical data mining has been prominent in the literature. This section gives a brief review of some basic concepts from prior works related to the development of our visual analysis of free-textual clinical diagnosis records.

Mining knowledge from a free-textual clinical diagnosis corpus has been pursued by many researchers. It was often referred to as text mining and required natural language processing to format unstructured information. For example, Heinze et al. [Heinze, Morsch et al. 2001] employed natural language processing to extract medical knowledge from a free-textual clinical diagnosis corpus about asthma, gallbladder disease and acute myocardial infarction. Mamlin et al. [Mamlin, Heinze et al. 2003] also described a natural language processing system to extract findings from radiology reports. Zhou et al. [Zhou, Han et al. 2006] presented a medical information extraction system to mine a variety of patient information from free-textual diagnosis records. The system used three major text mining components to perform the task of text classification. Other systems can be found in the existing literature [Mamlin, Heinze et al. 2003].

There is still much to explore with regards to presenting mining results visually. Exploring information from free-textual clinical diagnosis records with visualization is a necessary component of a complete and convenient data analysis environment. An alternative to the traditional text mining approach is to employ techniques such as visual clustering. Visual techniques offer advantages that cannot be found in traditional approaches. For example, visual clustering allows users to visually inspect the clustering process and result. There are several visual mining techniques, such as SOM [Kohonen 1995; Kohonen, Kaski et al. 2000], parallel coordinate [Inselberg and Dimsdale 1990], multi-dimensional scaling (MDS) [Oliveira and Levkowitz 2003], principle component analysis (PCA) [Schroeder and Eyre 2003], and others [Jain and Dubes 1988; Fayyad, Grinstein et al. 2002; Do and Poulet 2003; Liu, Wang et al. 2008].

In recent years, SOM has attracted much attention and has been extensively applied to exploratory data analysis. SOM projects high-dimensional data onto a low-dimensional space. Similar high-dimensional data tend to project to the same neuron or nearby neurons on the map. One of the most important properties of SOM is that it does not need a predetermined parameter for the total number of clusters. Determining this parameter tends to be a difficult task for other clustering algorithms.

Some studies from the existing literature have applied SOM-based approaches to analyze textual data. Lagus et al. [Lagus, Kaski et al. 2004] applied a SOM-based approach, WEBSOM, to organize a vast document collection according to textual similarities on a two-dimensional map. Martin-Valdivia and Garcia-Vega [Martin-Valdivia, Urena-Lopez et al. 2007] presented an LVQ algorithm that was a supervised version of the Kohonen model and performed pattern classification tasks. They applied their system to text categorization and word sense disambiguation. Lin [Lin 1997] developed a map display for information retrieval. The system was an unsupervised Kohonen model and was employed to visually analyze text documents. Merkl and Rauber [Merkl and Rauber 2000] applied SOM to clas-

sify text in a digital library environment. Other applications based on SOM can be found in references [Wiener, Pedersen et al. 1995; Chen, Houston et al. 1998; Dittenbach, Merkl et al. 2001].

# 3. Analysis of Clinical Diagnosis Records with ICD-9 Codes

## 3.1 Observation on a clinical diagnosis record

Clinical diagnosis records, which are often free-textual documents, consist of unstructured sentences and paragraphs. Data mining algorithms are usually suitable for structured data. The challenges of analyzing clinical diagnosis records with ICD-9 codes come from several factors, including unstructured data, poor data quality, imbalanced data, the limited vocabulary of the ICD-9 system, synonymous words, abbreviations and large similarities between disease subcategories.

We give a diagnosis example, shown in Fig. 1, that is written by a clinical doctor. The patient's condition is about a disease of the circulatory system. The ICD-9 code that contains this category of disease is "428.0." (The appendix provides more information about ICD-9 codes.) This diagnosis example suffers from common

病歷編號：12465
The 70 y/o male was diagnosed to have chronic renal failure with regular hemodialysis for 8 years. He denied any history of diabetes mellitus, hypertension or heart disease before. He was noted to have pulmonary TB with medical treatment for 6 months 20+ years ago. He was in his usual state of health until 1+ year before admission when he began to suffered from exertional dyspnea and abdominal fullness. No chest pain, orthopnea or proxysmal nocturnal dyspnea was experienced. The symptoms became worse in recent months. He came to our CV OPD for help. Mild dilated LV, LA and mild anterior-septal LV hypokinesis with impaired LV global function, thicken pericardium, moderate pulmonary hypertension and mild pericardiac effusion were noted by cardiac echo. Hepatomegaly, hepatic vein dilatation and ascites were also noted by abdominal echo. Severe abdominal fullness, severe jugular vein engorgement and Kussmaul's sign were also noted. Constrictive pericarditis was suspected. He was admitted for catheterization to comfirm the diagnosis and further work-up.
Social History
Smoking: (+) 1/2 PPD for 50+ years.
Alcohol: (-)
Betal nuts chewing: (-)
Past History
DM(-)
HTN(-)
Hyperlipidemia(-)
CRF in uremia stage with regular hemodialysis for 8 years.
Pulmonary TB(+) post meds. 20+ years ago
Asthma(-)
Peptic ulcer(-)
Denied other major medical illness
No major operation history
---
@ Risk factors for CAD:
DM(-),HTN(-),hypercholesterolemia(-),male(-),obesity(-), smoking(+),old age (+),family history(-),type A personality(+)
申報 ICD9 疾病碼：428.0

**Fig. 1** *A clinical diagnosis record is written by a clinical doctor. The ICD-9 code is assigned 428.0.*

spelling errors. Furthermore, abbreviations are also popularly used in these medical documents. The problem of synonymous words (in which distinct words have the same meaning) is a well-known and influential factor in the information retrieval field. Proper processes are needed to ensure data quality prior to applying mining algorithms.

### 3.2 Keyword extraction

The preprocess tasks output clean documents and converts the unstructured documents into structured data using the vector space model [Han and Kamber 2001; Ananiadou and Mcnaught 2006]. The number of unique words (terms) determines the size of the vector. Thus, a document is represented as a vector.

The value of an element in the vector represents the weight (the degree of importance) of the corresponding word in the document. There are many ways to determine the weight of a word. In this paper, we use the well-known TF-IDF (Term Frequency-Inverse Document Frequency) weighting method to index keywords, where TF-IDF is frequently used in the information retrieval and text mining field.

We use TF-IDF to weight every term in the vector space. The importance of a term increases proportionally based on the number of times that a word appears in a single document, but decreases proportionally based on the frequency of the word across different documents of the corpus. Specifically, this measure contains two factors: the term frequency (TF) and inverse document frequency (IDF). TF is the term frequency in a given document, and IDF is the inverse document frequency of the term. The TF equation is:

$$\mathbf{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{1}$$

where $n_{i,j}$ is the number of occurrences of the considered term in document $d_j$ and the denominator is the number of occurrences of all terms in document $d_j$. The IDF equation is:

$$\mathbf{idf}_i = \log \frac{(\#D)}{\#\{d_j : t_i \in d_j\}}, \tag{2}$$

where $\#D$ is total number of documents in the corpus and $d_j : t_i \in d_j$ is the number of documents where the term $t_i$ appears. The TF-IDF measure is

$$\mathbf{tfidf}_{i,j} = \mathbf{tf}_{i,j} \cdot \mathbf{idf}_i. \tag{3}$$

## 4. Visual Cluster Analysis of Clinical Diagnosis Records

We constructed systems with kernels based on SOM [Kohonen 1989] and ViSOM [Yin 2002; Yin 2002] to analyze free-textual clinical diagnosis records. We attempt to project dataspace distance onto a two-dimensional projection map so that the clustering structure of the data can be easily visualized. Fig. 2 shows the training procedure for diagnosis records using the SOM-based approach.

**Initialization**
0. Start with a SOM-based model
0. Initialize parameters needed for the SOM-based approach
0. Import clinical diagnosis records

**Begin**
1. Keyword extraction using Eq. (3)
2. Evaluate the distance using Eq. (4)
3. Train and update the current network with the SOM-based approach using Eq. (5) or Eq. (7)
4. Go to step 2 until the stop criteria are satisfied
5. Finish the training and output a 2-dimensional map

**Fig. 2** *The training procedure for clinical free-textual diagnosis records using the SOM-based approach.*

## 4.1 SOM approach

The SOM approach can nonlinearly project high-dimensional data onto a low-dimensional grid [Kohonen 1989]. The projection preserves the topological order from the input space; hence, similar data patterns in the input space will be assigned to the same map node or to nearby nodes on the trained map. First, the core process of the projection determines the best matching unit (BMU) from the map nodesfor each input pattern. The BMU is the node that is most similar to the input pattern. Then, the process updates the BMU and its neighborhood nodes to reduce the difference between those nodes and the input pattern. In short, the two key steps in the SOM training algorithm are to 1) determine the BMU and 2) update the BMU and its neighbors.

Specifically, an input pattern is a high-dimensional vector of real numbers, $\overrightarrow{x} = (x_1, \ldots, x_i, \ldots, x_n)$, in the Euclidean space, where $x_i$ is the value of the $i^{\text{th}}$ component. Each node on an associated SOM for an $n$-dimensional training dataset is also an $n$-dimensional real vector, $\overrightarrow{m_k} = (m_{1,k}, \ldots, m_{i,k}, \ldots, m_{n,k})$, where $m_{i,k}$ is the value of the $i^{\text{th}}$ component of the $k^{\text{th}}$ node in the map. A distance function is usually employed to measure similarity. Smaller distances indicate higher levels of similarity. Searching BMU for an input $\overrightarrow{x}$ is to search a $\overrightarrow{m_k}$ node such that the distance between $\overrightarrow{x}$ and $\overrightarrow{m_k}$ is smallest. A typical method for computing the distance $d(\overrightarrow{x}, \overrightarrow{m_k})$ uses the Euclidean distance function. The distance between $\overrightarrow{x}$ and $\overrightarrow{m_k}$ is calculated using Eq. (4).

$$d(\overrightarrow{x}, \overrightarrow{m_k}) = \| \overrightarrow{x} - \overrightarrow{m_k} \| = \left( \sum_{i=1}^{n} (x_i - m_{i,k})^2 \right)^{\frac{1}{2}} \tag{4}$$

Once the BMU is identified, the BMU and its neighbors are updated to reduce their differences with the input pattern. The update is centered at the BMU, and the adjustment amount decreases as distance from the BMU increases. Similarly, the update neighborhood also decreases with increasing training epochs. Formally, the update function for a neighborhood node $\overrightarrow{m_j}$ for giving $\overrightarrow{x}$ is Eq. (5).

$$\overrightarrow{m_j}(t+1) = \overrightarrow{m_j}(t) + \alpha(t) \times h_j(t) \times [\overrightarrow{x}(t) - \overrightarrow{m_j}(t)] \quad (5)$$

where $0 < \alpha(t) < 1$ is the learning-rate function, and $h_j(t)$ is the neighborhood function, which calculates the lattice distance between the BMU and $\overrightarrow{m_j}$. Both $\alpha(t)$ and the width of $h_j(t)$ decrease gradually with increasing step $t$.

### 4.2 ViSOM approach

Yin [Yin 2002; Yin 2002] notes that when SOM is used for visualization, the interneuron distances are not directly visible nor measurable on the map. The ViSOM model extends the SOM approach by alleviating the problem of data projection [Yin 2002]. To faithfully preserve the structure of the map's training data ViSOM considers not only the distance between two neurons (the winner and its neighbor) on the map but also their distance in the data space. Therefore, the updating force $[\overrightarrow{x} - \overrightarrow{m_j}]$ used in the SOM approach is rearranged and decomposed into two forces $[\overrightarrow{x} - \overrightarrow{m_v}]$ and $[\overrightarrow{m_v} - \overrightarrow{m_j}]$, where $\overrightarrow{m_v}$ is the BMU. The update function for the $j^{\text{th}}$ neuron for giving $\overrightarrow{x}$ is represented as Eq. (6)

$$F_j(\overrightarrow{x}) \equiv \overrightarrow{x} - \overrightarrow{m_j} = [\overrightarrow{x} - \overrightarrow{m_v}] + [\overrightarrow{m_v} - \overrightarrow{m_j}] \quad (6)$$

ViSOM takes into account the distance between the winner neuron $\overrightarrow{m_v}$ and its neighborhood neuron $\overrightarrow{m_j}$ both in the data space and on the map. The Eq. (6) is unobservable. Specifically, the update function for a neighborhood node $\overrightarrow{m_j}$ with respect to a winner neuron $\overrightarrow{m_v}$ for giving $\overrightarrow{x}$ is written as Eq. (7).

$$\overrightarrow{m_j}(t+1) = \overrightarrow{m_j}(t) + \alpha(t) \times h_{v,j}(t) \left( [\overrightarrow{x}(t) - \overrightarrow{m_v}(t)] + [\overrightarrow{m_v}(t) - \overrightarrow{m_j}(t)] \times \left( \frac{d_{v,j}}{\Delta_{v,j}} - 1 \right) \right) \quad (7)$$

where $h_{v,j}(t)$ is the lattice distance between the $\overrightarrow{m_v}$ and $\overrightarrow{m_j}$ at the $t^{\text{th}}$ iteration, $d_{v,j}$ is the distance between the weights of neurons $\overrightarrow{m_v}$ and $\overrightarrow{m_j}$. in the data space, $\Delta_{v,j}$ is the lattice distance between those neurons' locations on the map, and $\lambda$ is a positive pre-specified resolution parameter. Parameter $\lambda$ represents the desired interneuron distance reflected in the input space and depends on the size, data variance, and required resolution of the map. The smaller the value of $\lambda$ is, the higher resolution the map can be. We follow a previous study [Hsu and Lin 2012] to set $\lambda$ to 1.5.

## 5. Experiments

Visual representation is considered one of the most effective ways to disseminate information. To analyze clinical diagnosis records we constructed an interface that visually represents the trained ViSOM and SOM maps. All of the components in our system were developed using the programming tool C++ Builder 6.

In later sections, we briefly illustrate several interesting applications that could help clinical doctors. The view of two-dimensional ViSOM and SOM maps can effectively provide clinical doctors with an understanding of the contextual basis for

clinical diagnosis records. The analyses rely on performance comparisons between ViSOM and SOM basedon the two major aspects: 1) cluster analysis including and 2) interesting pattern discovery. First, we evaluate the performance and discuss the cluster analysis with satisfactory visual representation, approximate diagnosis records and quantitative evaluation of clustering. Second, we expect to discover interesting patterns for the clinical records projected in the map nodes.

## 5.1 Experimental data

The clinical diagnosis records were collected from May 2004 to August 2005 in a hospital in Taiwan. There were a total of 17,014 records. We ignored infrequent data; specifically, we ignored all ICD-9 codes that were recorded less than 50 times. We randomly selected 50 unique ICD-9 codes that were used in a total of 1,926 free-textual records. The size of the dimension is 6,128.

## 5.2 Visual Cluster Analysis

### 5.2.1 Visual representation for analyzing clinical diagnosis records

We demonstrated learning results on the topographic map for clinical diagnosis records using the ViSOM and SOM approaches. The sizes of the ViSOM and SOM maps were set to $15 \times 15$. The other parameterswere set according to recommendations in the literature [Yin 2002; Yin 2002; Hsu 2006].

Fig. 3 and Fig. 4 show the projected results for ViSOM and SOM, respectively. A map node with no projected record is displayed as a blank space; in contrast, a map node with projected records is displayed as a black dot. The dot is bigger when the amount of the projected records is larger. Each record has a single ICD-9 code. The ICD-9 codes are displayed on the corresponding map nodes.

We evaluated the visual performance of the ViSOM and SOM maps and performed cluster analysis (see Fig. 3 and Fig. 4). Specifically, we used the DBScan algorithm [Ester, Kriegel et al. 1996] to partition the map nodes into clusters with assigned labels. The results are shown in Fig. 5(a) and Fig. 5(b), respectively.

We observe that ViSOM and SOM could potentially preserve the topological relationship among ICD-9 codes. In general, data having the same ICD-9 code were often projected near each other, and data having different ICD-9 codes were often projected apart from each other in the visual map. For example, in the bottom-right corner in Fig. 5(a), we observe that the $18^{\text{th}}$ cluster consists of a total of seven map nodes. In this cluster, the diagnosis records for respiratory system diseases (ICD-9 codes: 460-519) are on six nodes, and the records with digestive system diseases (ICD-9 codes: 520-579) are on one node. Furthermore, we observe that the diagnosis records on the $19^{\text{th}}$ cluster are about respiratory diseases and digestive diseases, as shown in Fig. 5(b).

Comparing the outputs from the ViSOM and SOM approaches, the topographic results for ViSOM are more distinguishable than for SOM. In other words, the boundary between clusters for ViSOM is larger than for SOM. In the ViSOM output, the data between clusters could be separated by wider blank spaces. The boundary between the $14^{\text{th}}$ and $17^{\text{th}}$ clusters of the ViSOM output provides an

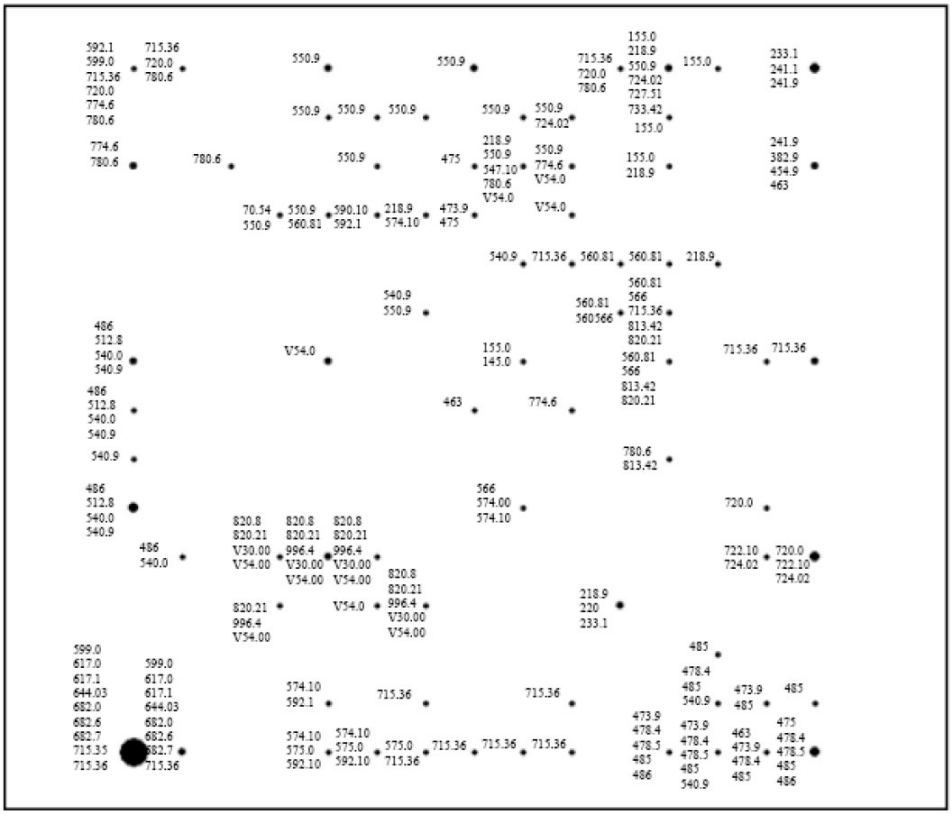


**Fig. 3** *Topographic results for SOM with the display of ICD-9 codes.*

example of wide spacing. The boundary between the 15$^{\text{th}}$ and 18$^{\text{th}}$ clusters also provides an example of wide spacing (see Fig. 5(a)). In contrast, because the SOM approach projects data nearly everywhere on the map of nodes (see Fig. 5(b)), the clusters are not as visually distinct.

An example of a practical application involves a clinical doctor who wants to assign an ICD-9 code to a new clinical diagnosis record with the help of ViSOM. After training his diagnosis record on ViSOM, the doctor's record is projected into the bottom-right of the map, as shown in Fig. 3. This visually suggests to the doctor that the ICD-9 code might be in the range of 463-486 (respiratory system diseases) rather than 540.9 (digestive system diseases). Furthermore, the map visually gives interesting information about trends between the diagnosis records and the distribution of records.

### 5.2.2 Approximate diagnosis records with topographic browser

The documents projected from high-dimensional space onto the viewable two-dimensional map nodes was useful for the task of document collection and search [Lagus, Kaski et al. 2004]. Furthermore, exploring diagnosis records on the map

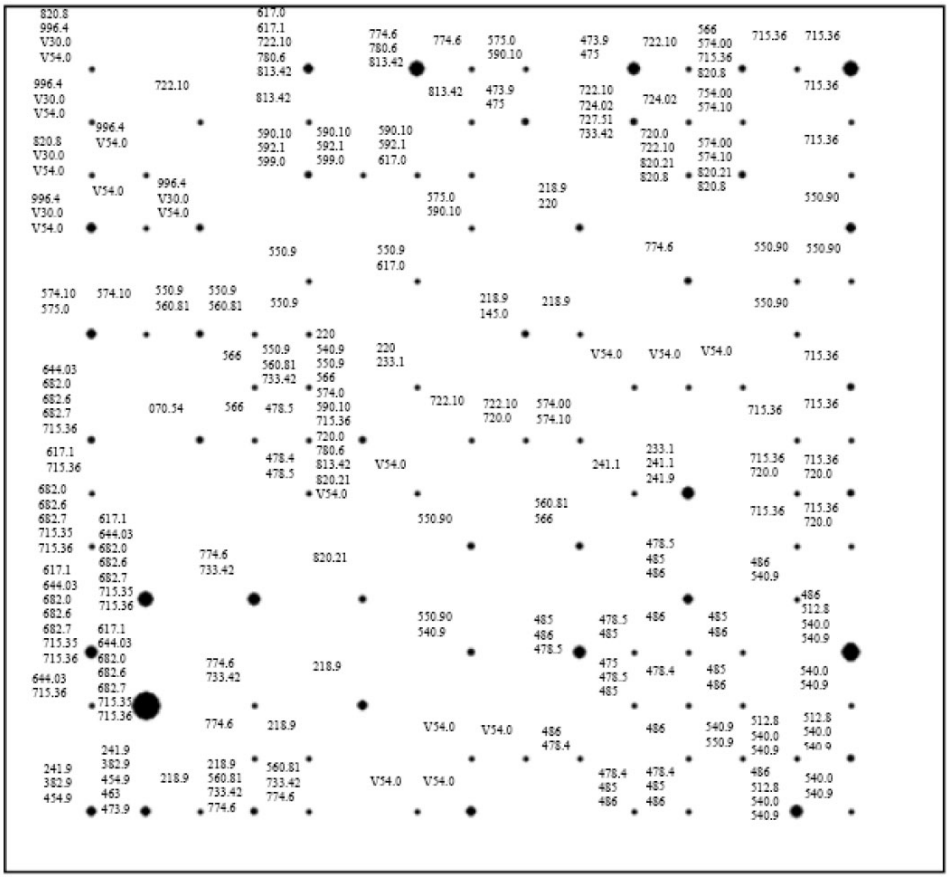
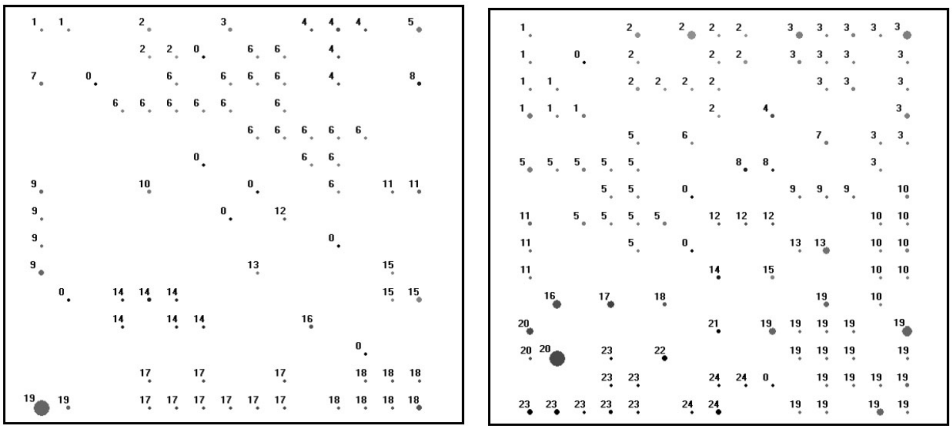


**Fig. 4** *Topographic results for SOM with the display of ICD-9 codes.*

nodes can support searches for more relevant records at the same nodes or the neighboring nodes because similar records in the data space are usually projected to the same neuron or nearby neurons in the visual map space. For example, the search for relevant records could start at a satisfactory map node and then extend to neighboring map nodes to retrieve relevant records. Therefore, to correctly approximate documents the map nodes could be adapted to support an effective document collection and search effort.

To demonstrate the effectiveness of approximating documents on the viewable map nodes we were interested in observing the distances in two aspects: 1) the geographic distance and 2) the prototype distance, where the distances were calculated using the Euclidean distance metric. First, we calculated the distance between the coordinates of nodes in the map. Each coordinate is a two-dimensional space. Second, we calculated the prototype distance, which represents the difference between the neurons with the original representation space. We then applied the U-matrix approach [Siemon 1990] to visually display the prototype distance in the map. For the U-matrix, the light area between neurons represents the small prototype

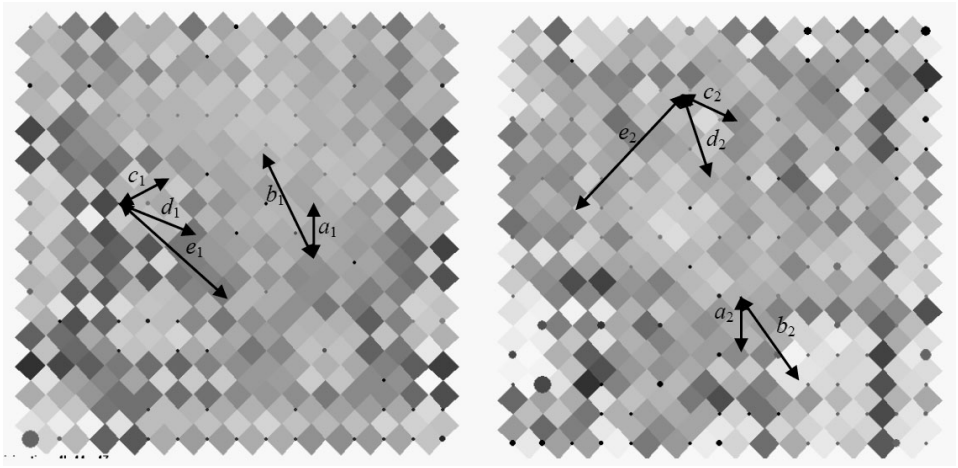


(a) clustering nodes for ViSOM, (b) clustering nodes for SOM

**Fig. 5** *Clustering for topographic nodes from (a) ViSOM and (b) SOM.*

distance, whereas the dark area between neurons represents the large prototype distance.

We briefly illustrated several geographic distances between nodes according to $a_1 < b_1$ and $c_1 < d_1 < e_1$ on the ViSOM map and $a_2 < b_2$ and $c_2 < d_2 < e_2$ on the SOM map. The cases are shown in Fig. 6. Furthermore, we examined whether the variations among prototype distances are similar to the variations among geographic distances. The prototype distances between nodes are listed in Tab. I.




(a) ViSOM map (b) SOM map

**Fig. 6** *Several geographic distances between nodes for (a) a ViSOM map with* $a_1 < b_1, c_1 < d_1 < e_1$ *and (b) a SOM map with* $a_2 < b_2, c_2 < e_2$.

| ViSOM | | SOM | |
|---|---|---|---|
| item | prototype distance | item | prototype distance |
| $a_1$ | 0.498 | $a_2$ | 1.096 |
| $b_1$ | 0.653 | $b_2$ | 0.968 |
| $c_1$ | 0.598 | $c_2$ | 0.911 |
| $d_1$ | 0.788 | $d_2$ | 0.965 |
| $e_1$ | 0.906 | $e_2$ | 0.694 |

**Tab. I** *Prototype distances evaluated from the arrows, listed in Fig. 6.*

We observe that the ViSOM approach keeps the map nodes ordered. The lengths of arrows in the ViSOM map are longer, and the prototype distances between the nodes are larger. For example, the prototype distance $a_1$ is smaller than $b_1$, and the prototype distances $c_1$ and $d_1$ are smaller than $e_1$ (see Tab. I). In contrast, the map nodes in the SOM map were easily disordered. The prototype distance of $e_2$ illustrates this: although the prototype distance of $e_2$ is much smaller than $c_2$ and $d_2$, the length of the arrow for $e_2$ is the longest of the three.

With respect to approximating clinical diagnoses, the ViSOM approach outperforms the SOM approach. The ViSOM approach approximates clinical diagnosis records better than the SOM approach.

### 5.2.3 Quantitative evaluation of clustering

In this section, we evaluated the quantitative performance of partitioning map nodes into clusters using clustering algorithms. The advantage of this analysis is that it relies on map nodes that can be visually partitioned into clusters, and the clustering results can easily be seen on the map. However, this algorithm cannot predetermine a specific number of clusters in advance. Alternatively, we used other clustering algorithms to partition nodes into clusters to handle the task of performance comparison.

The four clustering algorithms are $K$-means, self-organizing map (SOM), complete-link hierarchical clustering (HC) and partitioning around medoids (PAM). These algorithms were tested for cluster analysis on map nodes. For each algorithm, the number of clusters was set to $M \in \{2, 3, \ldots, 30\}$ to partition map nodes into $M$ clusters. With SOM, we followed the study referenced in [Vesanto and Alhoniemi 2000] to obtain a user-defined number of clusters for nodes.

The method to partition map nodes into clusters has two major steps. First, we carried the centers for all map nodes that had at least one diagnosis record projected onto it. The center was measured using the mean of data based on the corresponding node. Second, these centers were partitioned into clusters using clustering algorithms. The centers were partitioned together and represent the fact that data on the corresponding nodes were also partitioned together. Furthermore, we used the Davies-Bouldin index (DBI) [Davies and Bouldin 1979] to measure performance because DBI is frequently used for cluster evaluation. Lower DBI values indicate better clustering performance. Fig. 7 shows the DBI values for

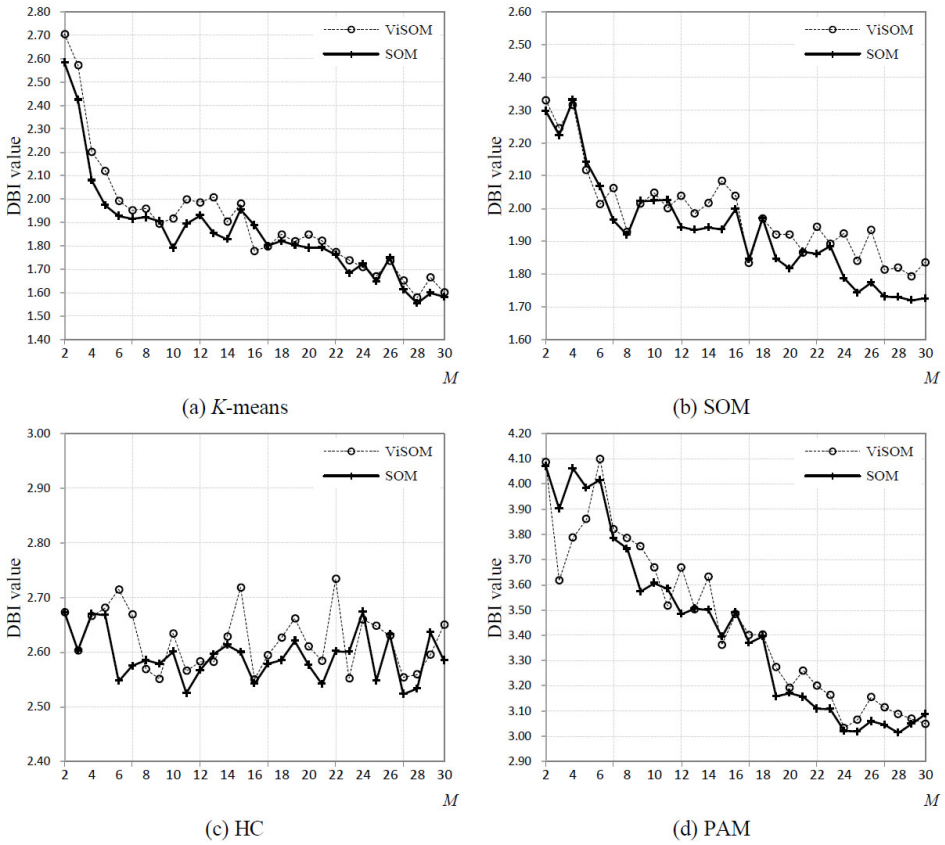



**Fig. 7** *DBI value according to M clusters using (a) K-means, (b) SOM, (c) HC and (d) PAM.*

various $M$ under the four clustering algorithms ($K$-means, SOM, HC, and PAM) used to partition map nodes into $M$ clusters.

Overall, SOM outperformed ViSOM. In Fig. 7(a), we found that the performance of SOM was better than ViSOM when $M < 15$, but was almost equivalent to ViSOM when $M \geq 15$. Fig. 7(b) indicates that the performance of SOM is better than ViSOM when $M$ is larger. Fig. 7(c) shows that SOM significantly outperforms ViSOM when $M = 7, 15$ and 22. Fig. 7(d) shows that the performance of ViSOM is comparable to the performance of SOM.

## 5.3 Interesting pattern discoveries

To correctly assign ICD-9 codes for diagnosis records is usually difficult; hence, the time-consuming process of manual assignments and frequent checks remains. There are two main errors that can occur with code assignment. First, diagnosis records that are similar are assigned different ICD-9 codes. Second, records that are

different are assigned the same ICD-9 code. Therefore, overcoming these difficulties is straightforward. We expect to discover interesting patterns in the map nodes that can contribute to overcoming some of these difficulties. Because the authors of this paper are not members of the medical profession, they could not provide an adequate analysis of clinical diagnosis records with ICD-9 codes. Dr. Wu and Dr. Chang of the National Taiwan University Hospital's Yun-Lin Branch shared their domain knowledge and provided us with useful explanations.

Our analyses relied on the visual view, which is different from the analyses using nonvisual approaches. The diagnosis records with the same keywords (i.e., the same symptoms of illness) were projected onto the same map node or onto neighboring nodes. The records with different keywords were projected apart from each other. Because the projected map nodes visually show relationships between topographic map nodes, they could be helpful in discovering interesting patterns. Therefore, we examine which similar records had different ICD-9 codes and which dissimilar records had the same ICD-9 code from map nodes.

### 5.3.1 Similar records with different ICD-9 codes

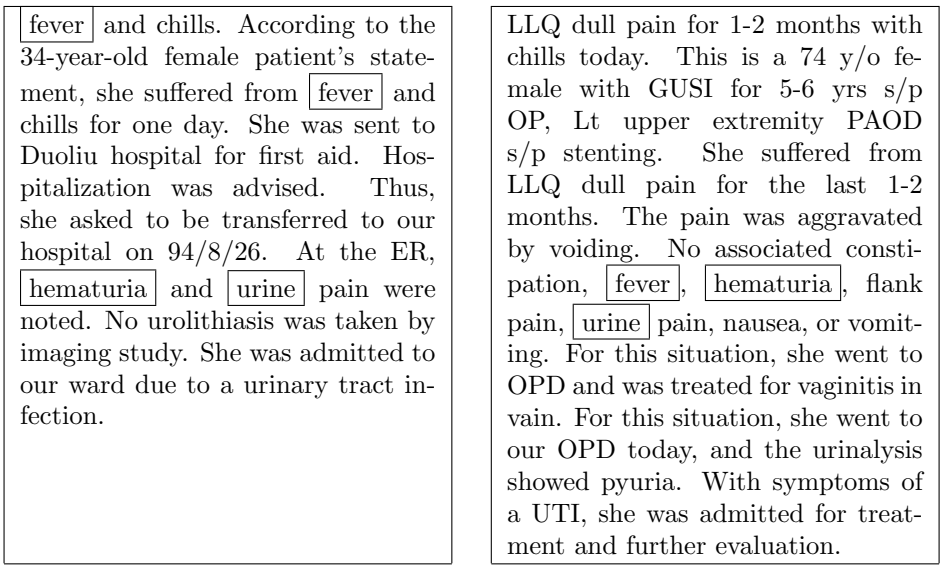
An ICD-9 code is used to describe a unique symptom of illness; however, a symptom might result in several different ICD-9 codes. For example, consider a woman who has abdominal or urinary pain. She can visit either the obstetrics and gynecology department or the urology department. In practice, the doctors from each department might evaluate the patient differently, give different medical advice or prescribe different medications. Hence, this diagnosis record is assigned different ICD-9 codes. Because these records were projected on the map nodes, the nodes could effectively allow us to detect their similarities.

Recall that Fig. 5(a) shows the clustering result of the topographic map. We observe that the $19^{\text{th}}$ cluster has two map nodes where the diagnosis records had different ICD-9 codes. To further discuss the task of discovering interesting patterns we randomly selected two diagnosis records from the two map nodes. These two records are shown in Fig. 7(a) and Fig. 7(b). We also listed the important keywords (i.e., large vector values of corresponding nodes) that simultaneously occurred in both records and displayed them in the rectangular frame.

In Fig. 8, the terms "fever", "hematuria" and "urine" were interesting illnesses that clinical doctors assigned ICD-9 codes to at the time of writing the discharge summary or the clinical medical diagnosis. When a new diagnosis record was created and needed to be assigned to an ICD-9 code, these terms provided valuable references for the code assignment. We also listed two examples, as shown in Fig. 9 and Fig. 10. The terms "knee" and "leukocytosis" might also be the interesting patterns.

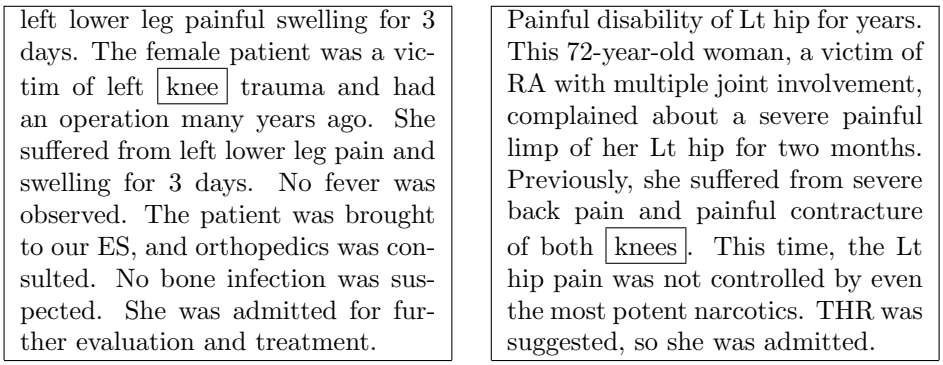
### 5.3.2 Dissimilar records with the same ICD-9 code

Many different terms can describe an ICD-9 code; thus, dissimilar records might have the same ICD-9 code. Furthermore, clinical doctors have different cognitions and habits of writing an anamnesis. To correctly assign ICD-9 codes for these dissimilar records human experts usually need to manually assign the code. Therefore,



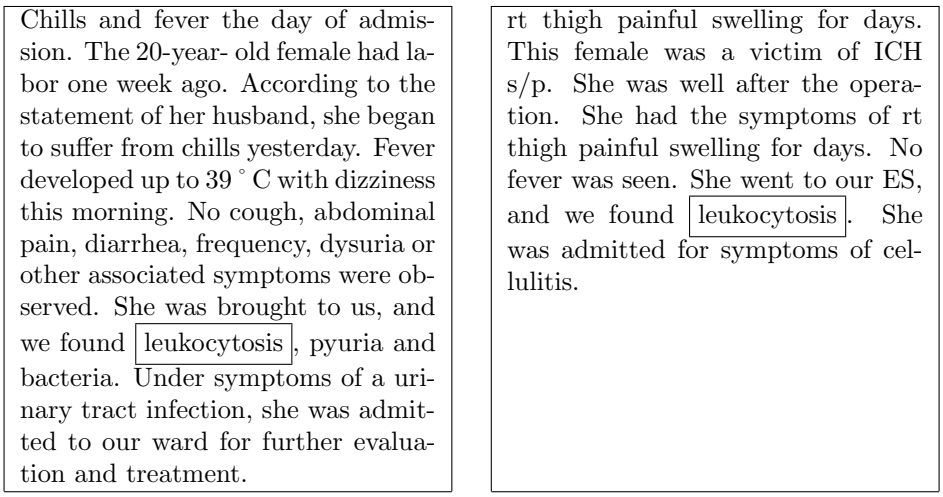

(a) ICD-9 code is 599.0 (b) ICD-9 code is 644.03

**Fig. 8** *The ICD-9 codes are (a) 599.0, which is about diseases of the genitourinary system, and (b) 644.03, which is about complications of pregnancy, childbirth, and puerperium. The records were written by the clinical doctors who worked in the urological department and the obstetrics and gynecology department, respectively.*




(a) ICD-9 code is 682.6 (b) ICD-9 code is 715.35

**Fig. 9** *The ICD-9 codes are (a) 682.6, which is about diseases of the skin and subcutaneous tissue, and (b) 715.35, which is about diseases of the musculoskeletal system and connective tissue. The records were written by clinical doctors who worked in the plastic surgery department and the orthopedics department, respectively.*




(a) ICD-9 code is 599.0 (b) ICD-9 code is 682.6

**Fig. 10** *The ICD-9 codes are (a) 599.0, which is about diseases of the genitourinary system, and (b) 682.6, which is about diseases of the skin and subcutaneous tissue. The records were written by the clinical doctors who worked in the kidney department and the plastic surgery department, respectively.*

we tested whether the projected map nodes could effectively recognize dissimilar records with the same ICD-9 code.

We randomly selected two dissimilar diagnosis records from different map nodes. The map nodes were not close together, implying that the contents of the records were not similar. These two records are shown in Fig. 11. We observe that the doctors used different vocabularies to represent the same symptom of illness. For example, the term "fever" was used in Fig. 11(a), while the sentence "temperature 38.2$^{o}$ (Celsius) by ear was found" was used in Fig. 11(b). This finding could be used to design critical strategies and linguistic conventions for assigning correct ICD-9 codes effectively.

# 6. Discussions and Conclusions

In this paper, we illustrated several interesting applications that can be used to help clinical doctors analyze real diagnosis records with ICD-9 codes. The analyses compared performance between the ViSOM and SOM approaches based on the following aspects: (1) visual cluster analysis for analyzing clinical diagnosis records; (2) approximate diagnosis records with topographic browser; (3) interesting pattern discoveries; and (4) quantitative evaluation of clustering. The qualitative evaluation from the prior three aspects shows that ViSOM provides better visual perception than SOM in projecting high-dimensional diagnosis records into a two-dimensional map. On the other hand, the quantitative evaluation shows that SOM outperforms ViSOM in partitioning map nodes into clusters using four clustering

| (a) | (b) |
|---|---|
| Fever was noted today. The 3-month-old male infant had good health in the past. Cough and rhinorrhea were noted for about one month. She has a fever. She was brought to LMD and our department for help. The above symptoms did not improve. Decreased activity and poor intake developed for 2 days. She visited our department for help again. Under the symptom of fever, she was admitted to our department for further evaluation. | Nasal obstruction, cough with sputum for 3 days. The 2-month-old infant had been in good health since birth. However, she suffers from nasal obstruction and her throat has sputum for the last 3 days. She was brought to our OPD for help, where we found her temperature to be 38.2 °C by ear.. Her appetite and activity have decreased. She does not have diarrhea or vomiting. Under the impression of bronchopneumonia, she was admitted to our NOR for further evaluation and treatment. |

**Fig. 11** *The ICD-9 code for both records is 485, which is about diseases of the respiratory system. The records were written by clinical doctors who worked in the pediatrics department.*

algorithms ($K$-means, SOM, HC and PAM). ViSOM does not perform as well as SOM in the cluster analysis for map nodes. More importantly, however, ViSOM is *comparable* to SOM because it provides a satisfying visual scenario for analyzing diagnosis records.

Because the basic SOM approach has been continuously used for various applications, as described in the literature, we discussed the characteristics of ViSOM and SOM through the analysis of real diagnosis records with ICD-9 codes. ViSOM and SOM are helpful for real diagnosis records and provide valuable visual analyses that are not provided by traditional, nonvisual approaches. Furthermore, our analysis can be used to support organizational decision-making activities in the future.

## Acknowledgments

The authors would like to thank the reviewers of this article for their valuable suggestions. We also thank the medical doctors Dr. Wu and Dr. Chang of the National Taiwan University Hospital's Yun-Lin Branch for providing their explanations of clinical diagnosis records with ICD-9 codes. This research was supported by the National Science Council, Taiwan, under grant NSC 99-2410-H-146-001-MY2.

# References

[1] Ananiadou S., Mcnaught J.: Text Mining for Biology And Biomedicine, Artech House, Inc., 2006.

[2] Boeckmann B., Bairoch A., et al.: The SWISS-PROT protein knowledgebase and its supplement TrEMBL. Nucleic Acids Res., **31**, 2003, pp. 365–370.

[3] Cerrito P. B., Cerrito J. C.: Data and text mining the electronic medical record to improve care and to lower costs. Proceedings of the Thirty-first Annual SAS User Group International Club, San Francisco, 2006.

[4] Chen H., Houston A., et al.: Internet browsing and searching: User evaluations of category map and concept space techniques. Journal of the American Society for Information Science, **49**, 1998, pp. 582–603.

[5] Davies D. L., Bouldin D. W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, **1**, 4, 1979, pp. 224–227.

[6] Dittenbach M., Merkl D., et al.: Hierarchical clustering of document archives with the growing hierarchical self-organizing map. In: Proceedings of international conference on artificial neural networks, 2001.

[7] Do T.-N., Poulet F.: Incremental SVM and Visualization Tools for Bio-medical Data Mining. Proceedings of the European Workshop on Data Mining and Text Mining for Bioinformatics, 2003.

[8] Ester M., Kriegel H. P., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996.

[9] Fayyad U. M., Grinstein G., et al., Eds.: Information Visualisation in Data Mining and Knowledge Discovery. San Francisco, Morgan Kaufmann Publishers, 2002.

[10] Franz P., Zaiss A., et al.: Automated coding of diagnoses: Three methods compared. Proc. AMIA Symp., 2000.

[11] Han J., Kamber M.: Data mining concepts and techniques, San Francisco: Morgan Kaufmann, 2001.

[12] Heinze D. T., Morsch M. L., et al.: Mining Free-text Medical Records. Proc. AMIA Symp., 2001.

[13] Hsu C.-C., Lin S.-H.: Visualized analysis of mixed numeric and categorical data via extended self-organizing map. IEEE Transactions on Neural Networks and Learning Systems, **23**, 1, 2012, pp. 72-86.

[14] Hsu C. C.: Generalizing self-organizing map for categorical data. IEEE Transactions on Neural Networks, **17**, 2, 2006, pp. 194-204.

[15] Inselberg A., Dimsdale B.: Parallel Coordinates: A Tool for Visualizing Multi-dimensional Geometry. IEEE Visualization, 1990, pp. 361-378.

[16] Jain K., Dubes R. C.: Algorithms for clustering data, Prentice Hall, New Jersey, 1988.

[17] Kohonen T.: Self-Organization and Associative Memory, Springer-Verlag, Berlin-Heidelberg-New York-Tokyo, 3rd edition, 1989.

[18] Kohonen T.: Self-organization and associative memory. Berlin: Springer-Verlag, 1995.

[19] Kohonen T., Kaski S., et al.: Self organization of a massive document collection. IEEE Transactions on Neural Networks, 11, 2000, pp. 574-585.

[20] Lagus K., Kaski S., et al.: Mining massive document collections by the WEBSOM method. Information Sciences, **163**, 2004, pp. 135-156.

[21] Larkey L. S., Croft W. B.: Combining Classifiers in Text Categorization. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1996.

[22] Lin X.: Maps displays for information retrieval. Journal of the American Society for Information Science, **48**, 1997, pp. 40-54.

[23] Liu Y., Wang X., et al.: ConSOM: a conceptional self-organizing map model for text clustering. Neurocomputing, **71**, 2008, pp. 857-862.

[24] Mamlin B. W., Heinze D. T., et al.: Automated Extraction and Normalization of Findings from Cancer-Related Free-Text Radiology Reports. AMIA Annu. Symp. Proc., 2003.

[25] Martin-Valdivia M. T., Urena-Lopez L. A., et al.: The learning vector quantization algorithm applied to automatic text classification tasks. Neural Networks, **20**, 6, 2007, pp. 748-756.

[26] Merkl D., Rauber A.: Digital libraries classification and visualization techniques. In: Proceedings of international conference on digital libraries: Research and practice, 2000.

[27] Oliveira M. C. F. D., Levkowitz H.: From Visual Data Exploration to Visual Data Mining: A Survey. IEEE Transactions on Visualization and Computer Graphics, **9**, 3, 2003, pp. 378-394.

[28] Schroeder M., Eyre C.: Visualisation and Analysis of Bibliographic Networks in the Biomedical Literature: A Case Study. Proceedings of the European Workshop on Data Mining and Text Mining for Bioinformatics, 2003.

[29] Schuler G. D.: Entrez: molecular biology database and retrieval system. Methods Enzymol, **266**, 1996, pp. 141-162.

[30] Siemon A. U. a. H. P.: Kohonen's self organizing feature maps for exploratory data analysis. In: Proceedings of International Neural Networks Conference (INNC), Dordrecht, Netherlands, 1990.

[31] Soualmia L. F., Darmoni S. J.: Combining different standards and different approaches for health information retrieval in a quality-controlled gateway. International Journal of Medical Informatics, **74**, 2005, pp. 141-150.

[32] Vesanto J., Alhoniemi E.: Clustering of the self-organizing map. IEEE Transactions on Neural Networks, **11**, 3, 2000, pp. 586-600.

[33] Wiener E., Pedersen J., et al.: A neural network approach to topic spotting. In: Proceedings of the 4th annual symposium on document analysis and information retrieval, 1995.

[34] Wilbur W. J., Coffee L.: The effectiveness of document neighboring in search enhancement. Information processing and management, **30**, 1994, pp. 253-266.

[35] Yin H.: Data visualization and manifold mapping using the ViSOM. Neural Networks, **15**, 2002, pp. 1005-1016.

[36] Yin H.: ViSOM – a novel method for multivariate data projection and structure visualization. IEEE Transactions on Neural Networks, **13**, 1, 2002, pp. 237-243.

[37] Zhou X., Han H., et al.: Approaches to text mining for clinical medical records. Proceedings of the 2006 ACM symposium on Applied computing, 2006.

# 7. Appendix

International Classification of Diseases (ICD) is proposed and maintained by World Health Organization with the intent to standardize the coding of diseases. There are seventeen main categories of diseases and three supplementary categories of relevant information of diseases as shown in Tab. II. The ICD codes form a hierarchical structure which consists of the section, grouping, category and subcategory.

| Disease name | ICD-9 code |
|---|---|
| Infectious and Parasitic Diseases | 001-139 |
| Neoplasms | 140-239 |
| Endocrine, Nutritional and Metabolic Diseases, and Immunity Disorders | 240-279 |
| Diseases of the Blood and Blood-forming Organs | 280-289 |
| Mental Disorders | 290-319 |
| Diseases of the Nervous System and Sense Organs | 320-389 |
| Diseases of the Circulatory System | 390-459 |
| Diseases of the Respiratory System | 460-519 |
| Diseases of the Digestive System | 520-579 |
| Diseases of the Genitourinary System | 580-629 |
| Complications of Pregnancy, Childbirth, and the Puerperium | 630-677 |
| Diseases of the Skin and Subcutaneous Tissue | 680-709 |
| Diseases of the Musculoskeletal System and Connective Tissue | 710-739 |
| Congenital Anomalies | 740-759 |
| Certain Conditions Originating in the Perinatal Period | 760-779 |
| Symptoms, Signs, and Ill-Defined Conditions | 780-799 |
| Injury and Poisoning | 800-999 |
| V-code Supplementary Classification of Factors Influencing Health Status and Contact with Health Services | V01-V82 |
| E-code Supplementary Classification of External Causes of Injury and Poisoning | E800-E999 |
| M-code Morphology of Neoplasms | M8000/0-M9970/1 |

**Tab. II** *Diseases and injuries tabular index.*

The category of disease consists of a hierarchical structure. For example, Tab. III shows that code 001 for *Cholera* belongs to I.1 for *Intestinal Infectious Disease*, I.1 is part of I for *Infectious and Parasitic Diseases*, and code 001 can be extended to 001.0, 001.1 and 001.9 by different germs.

| | |
|---|---|
| I | Infectious and Parasitic Diseases |
| I.1 | Intestinal Infectious Disease |
| | |
| 001 | Cholera |
| 001.0 | Cholera due to Vibrio cholerate |
| 001.1 | Cholera due to Vibrio cholerate el tor |
| ... | ... |
| 001.9 | Cholera, unspecified |

**Tab. III** *An example of the ICD-9 hierarchical structure.*