Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 850–858
Expert Systems with Applications
www.elsevier.com/locate/eswa

Using GHSOM to construct legal maps for Taiwan's securities and futures markets

Jen-Ying Shih [a,*], Yu-Jung Chang [b], Wun-Hwa Chen [c]

[a] Department of Business Administration, Chang Gung University, 259, Wen-Hwa 1st Road, Taoyuang 333, Taiwan, ROC
[b] Institute of Information Science, Academia Sinica, Taipei, Taiwan
[c] Graduate Institute of Business Administration, National Taiwan University, Taipei, Taiwan, ROC

Abstract

A good legal knowledge representation system, capable of effectively providing investors with comprehensive legal knowledge, is needed for investors to prevent erratic behavior before investment decisions. This is especially important in Taiwan's securities and futures markets because the majority of market participants are individual investors who have limited access to legal knowledge about markets. Besides, the construction of the knowledge representation has to be automatic in order to efficiently handle the fast-growing and changeable legal information. Thus, we use the GHSOM algorithm to present a content-based and easy-to-use map hierarchy for Chinese legal documents in the securities and futures markets in the Chinese language. Meanwhile, an enhanced topic selection module and a web-based user interface are also proposed. The maps can be browsed on the web site (http://synteny.iis.sinica.edu.tw/legalmap/). To evaluate the legal maps, we apply two approaches, namely a validity test and a task experiment.

Keywords: Legal informatics; Knowledge representation; Self-organizing maps; Securities markets

1. Introduction

In legal informatics, successful searching of huge legal digital libraries depends a great deal on the user's ability to master the legal terminology (Schweighofer, Rauber, & Dittenbach, 2001). Due to cost considerations, traditional approaches to legal knowledge representation based on thesauri and classification by professional persons have gradually developed into semi-automatic or automatic approaches. Thus, how to automatically categorize, index, organize, present, and summarize the enormous amounts of legal documents to enable quick and efficient retrieval of accurate legal information is a key issue (Merkl & Schweighofer, 1997; Schweighofer et al., 2001; Schweighofer, Winiwarter, & Merkl, 1995; Thompson, 2001).

Most previous research in legal informatics has focused on Western languages – seldom on oriental languages, such as Chinese, Japanese, or Korean. In this paper, we study the effects of applying legal informatics to legal documents written in one of the above oriental languages by using an unsupervised learning algorithm called the growing hierarchical self-organizing map (GHSOM). As this algorithm has a proven performance record in Western language legal knowledge systems, we apply it to the construction of Chinese legal maps.

According to Schweighofer et al. (2001), it is better to segment the topics in legal documents when presenting legal knowledge. We, therefore, selected a very important topic, namely, legal knowledge of Taiwan's securities and futures markets, as our test environment, because – according to statistics published by the ROC government[1] – the

* Corresponding author. Tel.: +886 3 2118800x5410; fax: +886 2 27089989.
E-mail address: jyshih@mail.cgu.edu.tw (J.-Y. Shih).

[1] According to Major Indicators of the Securities and Futures Markets, Taiwan District, ROC, in 2005, the trading value of domestic individual investors is up to 74.6% of the total trading value, and 2.94 times larger than that of institutional investors.

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.10.031

majority of participants in these markets are individual investors. In contrast, institutional investors are the key players in the American and European markets. To protect individual investors, who are not as sophisticated as their institutional counterparts in gathering market information, Taiwan's securities and futures markets are highly regulated by several government agencies and the markets' self-regulating bodies, including the Securities and Futures Bureau (SFB), the Taiwan Securities Exchange (TSE), the ROC. Over-the-counter Securities Exchange (OTC), and Taiwan Futures Exchange (Taifex). However, as most investors are not familiar with the large number of laws, rules, and regulations promulgated by the competent authorities, it is hard for them to understand the relationships among the different types of legal information. As a result, they sometimes violate laws, and inadvertently commit crimes.

Currently, legal information retrieval systems used in Taiwan's markets emphasize key word search functions, but the search results are not presented in a meaningful way. As most market participants are not experienced in retrieving legal information, the retrieval procedure should be uncomplicated and user friendly so that legal knowledge about Taiwan's securities and futures markets can be accessed easily. Hence, the goal of this paper is to provide helpful and cost-effective legal guidance for market participants, including institutional and individual investors, employees of public companies, and securities and futures service providers.

The remainder of this paper is organized as follows. In Section 2, we introduce related research about legal knowledge retrieval. In Section 3, we review the self-organizing map (SOM) related literature and point out some drawbacks of applying growing hierarchical SOM (GHSOM) to legal maps. We then describe how to construct legal maps for Taiwan's securities and futures markets in Section 4, and present the legal maps in Section 5. In Section 6, we evaluate GHSOM by a validity test and a task experiment. Finally, in Section 7, we present our conclusions.

2. Literature review of legal knowledge systems

In recent years, many researchers have focused on the area of legal knowledge presentation. Their methods can be categorized as either supervised indexing approaches, or unsupervised learning models. The former use thesauri, ontology, or other classifications for a simple form of conceptual search. Some researchers apply knowledge-based models composed from thesauri and rules, such as the FLEXICON project (Smith et al., 1995). Case-based reasoning (CBR) plays an important role in developing some legal expert systems. In the SPIRE project, it is applied

Table 1
Summary of related research

Research	Methodology	Application
Flexicon project (Smith et al., 1995)	Knowledge base	A term extraction module that recognizes concepts, case citations, statute citations, and fact phrases resulting in a very helpful structured document profile. This profile is transformed into a weighted vector to enable a search for related documents or problems
SPIRE project (Daniels & Rissland, 1995)	CBR	Inference net presentations of important cases are used to search for similar documents
Salomon project (Moens, 1999)	CBR	Linguistic representations capable of abstracting legal documents automatically
Anandanpillai and Barta (1999)	CBR	The system, composed of the case knowledge base, pattern detection mechanism, case selection mechanism, adaptation mechanism and weight adjustment mechanism, is developed for housing discrimination law
ASHSD-III (Pal, 1999)	RBR & CBR	The goal of this research is to formulate a model by which one can find the appropriate precedents in an automated system. The ASHSD-III prototype has achieved the goal by using a model of case similarity assessment based on fuzzy proximity relations
BanXupport (Elhadi, 2000)	IR and CBR	The Statutes-Based Case Storage and Retrieval System (BanXupport) is a prototype system of case-based indexing and retrieval that is designed to represent and retrieve cases in bankruptcy law. It can assist lawyers in bankruptcy law and similar fields in doing their law research and reasoning using previously decided cases to solve new ones
KONTERM (Merkl et al., 1995)	SOM	Kobonen's self-organization feature map is used to represent 41 neutrally related text documents extracted from documents contained in the European Community law database, CELEX. The database prepares document weight vectors in the [0, 1] format, which represents every term (feature) in a document, whether or not such terms actually appear
KONTERM (Merkl & Schweighofer, 1997)	GSOM	The researchers present the results of a case study in legal document classification based on an experimental document archive comprised of 100 important treaties in public international law (5 MB of text). They use a layered architecture consisting of mutually independent unsupervised, neural networks. The length of each individual feature vector is equal to 1625 components
Schweighofer et al. (2001)	GHSOM & LabelSOM	Schweighofer et al. (2001) show the feasibility of using GHSOM and LabelSOM techniques in legal research by tests with text corpora in European case law. The results show the generalities and specialties of legal text corpora. The segmentation of a document improves the quality of labeling significantly. The next challenge will be a change from $tf \times idf$ vector representation to a modified vector representation that takes into account thesauri or ontologies considering learned properties of legal text corpora

to search for similar documents by inference net presentation of important cases (Daniels & Rissland, 1995). Anandanpillai and Barta (1999) develop a CBR system for housing discrimination laws. Pal (1999) builds a hybrid RBR–CBR knowledge-based system for legal decision-making processes. Elhadi (2000) presents an integrated application of CBR with an automatic indexing information retrieval (IR) component in the legal domain of bankruptcy law. In addition, several researchers have tried to generate abstracts of legal documents automatically in linguistic presentations (Moens, 1999). With regard to unsupervised learning models, in the KONTERM project (Schweighofer & Winiwarter, 1995) various documents are presented as feature vectors of terms extracted from the full text of a document, the context-sensitive rules, and the meta rules associated with the document. Improved SOM models are applied to represent legal knowledge (Merkl & Schweighofer, 1997, 2001). The above research is summarized in Table 1.

3. Literature review of SOMs

In this paper, we integrate the GHSOM and LabelSOM algorithms to construct legal maps. Since both algorithms are based on SOM, the SOM algorithm and its limitations in knowledge representation are introduced first. Then, the evolution of SOM algorithms, growing SOM (GSOM), hierarchical SOM (HSOM), and GHSOM are described in turn. We also point out some limitations of applying GHSOM to legal maps.

3.1. SOM

Kohonen (1982) proposed the SOM algorithm for clustering applications. The SOM model is widely used for generating topology-preserving maps and for data visualization. The remarkable characteristic of the SOM algorithm is that the similarities of the input data are mirrored to a very large extent by their geographical vicinity within the representation space. The SOM algorithm is shown in Fig. 1. This learning procedure leads to a topologically ordered mapping of the presented input signals. Similar types of input data are assigned to neighboring regions on the map (Kohonen, 1995). However, the decision about the best size of the various maps, as well as about the depth of the hierarchy, remains a non-trivial problem requiring some insight into the structure of the underlying document archive.

3.2. Improvements of SOMs

The SOM algorithm has three major drawbacks. The first is that the topology has to be fixed in advance, but this is not generally possible because the statistical properties of the given data are not always available. In order to avoid a tedious trial-and-error process to find the best size of map for a problem domain, several researchers have designed algorithms (GSOM) that automatically define the size of map (Fritzke, 1995; Rodrigues & Almeida, 1990). The network automatically chooses a height/width ratio suitable for the data distribution at hand. Locally accumulated statistical values are then used to determine where to insert

Begin

Set initial neighborhood parameters, learning rate parameters, and accepted error measurement.

Initialize weight vectors of the output layer

Repeat

For each input vector $x(t)$, t represents the time-stamp of the current learning iteration.

For each node in the output layer, calculate the distances between the weight vectors and the input vector: $D_i(t) = \|x(t) - m_i(t)\|$,

where $m_i(t)$ refers to the weight vector of unit i, and

$\|\cdot\|$ represents the Euclidean vector norm.

Determine the winner c: $c(t) : \|x(t) - m_c(t)\| = \min_i \{\|x(t) - m_i(t)\|\}$

Adapt the weight vectors according to the following learning rule:

$m_i(t+1) = m_i(t) + \alpha(t) \cdot h_{ci}(t) \cdot [x(t) - m_i(t)]$,

where $\alpha(t)$ represents a time-varying learning-rate decreasing over time, and

$h_{ci}(t)$ is a time-varying neighborhood function that takes into account the distance between the winner c and unit i $\|r_c - r_i\|$ within the output space.

The most often used function is $h_{ci}(t) = \exp\left(-\frac{\|r_c - r_i\|^2}{2 \cdot \delta(t)^2}\right)$.

Reduce the learning rate and radius of the neighborhood function

Until (Convergence or the maximum number of iterations is achieved)

End

Fig. 1. The SOM algorithm.

rows or columns of units whose weight vectors are the averages of nearby units. The growth process continues until the level of input signals falls below a bound for each unit in the network. The grid may then adapt its height/width ratio to the given pattern distribution.

The second drawback of applying SOM to document management is that it can only present similarities among the input data by a distance concept, not by a parent–child relationship. However, as knowledge presentation often relies heavily on hierarchical relationships, hierarchical SOM (HSOM) models have been proposed to resolve the problem. Generally, the training of a hierarchical feature map is performed sequentially from the first layer, i.e., one self-organizing map, down through the hierarchy. The maps of each layer are trained according to the standard self-organizing map learning process. As soon as the first layer map has reached a stable state, training continues with the maps of the second layer. In the second layer, each map is only trained with the input data assigned to the corresponding unit of the first layer map. Moreover, the length of the input vectors may be reduced by omitting those vector components that are equal to the original input vectors. Such an omission is possible because the corresponding features of the input vector are already represented in the previous level. As the size of the input vectors is reduced, the time needed to train the maps is also reduced. The training of the second layer is completed when each map has reached a stable state. The same training procedure is utilized to train the third and any subsequent layers of self-organizing maps (Merkl & Schweighofer, 1997).

The third drawback of the SOM model is that it lacks the interpretability of a trained SOM, i.e., a description of the granularities (topics) in maps. Reading and interpreting the structure and characteristics of a map display learned during the training process is difficult without expensive manual intervention. However, reading and interpreting are very important in knowledge representation and retrieval. This has led to the development of a number of enhanced visualization techniques to support the interpretation of self-organizing maps. Rauber et al. (1999) presented a novel LabelSOM approach for the automatic labeling of trained self-organizing maps based on the information provided by a trained SOM. Every unit of the map is labeled with the features that best characterize all the data mapped in that particular unit. This is achieved by using a combination of the quantization error of every feature and the relative importance of that feature in the weight vector of the unit (Rauber et al., 1999). Ideally, we need a method that can automatically label the units of a SOM based on the characteristics learned during the training process. In some situations, pre-classified information is available, such as in the WebSOM project (Honkela, Kaski, Lagus, & Kohonen, 1997), where the units of a SOM representing Usenet newsgroup articles are labeled with the name of the newsgroup or the newsgroup hierarchy that the majority of articles in a unit come from. This allows a kind of automatic assignment of labels to the units of a SOM by using the additional knowledge provided by the pre-classification of articles in newsgroups. Without a priori knowledge of the data, even information about the cluster boundaries does not reveal information about the relevance of a single attribute to the clustering and classification process. Instead, the labels help to identify the most important features within every unit and thus help a user understand the information presented by a particular unit. They can serve as the basis for simplified semi-automatic creation of class labels by allowing the user to choose the most appropriate terms from the automatically created list.

To overcome these limitations of SOMs within a uniform framework, Dittenbach, Rauber, and Merkl (2002) proposed an artificial neural network architecture, called GHSOM. It uses a hierarchical structure of multiple layers, where each layer consists of a number of independent SOMs. One SOM is used at the first layer of the hierarchy. For every unit on a map, a SOM might be added to the next layer of the hierarchy if the deviation of input data mapped to it reaches a predefined threshold. The principle is repeated with the third and any subsequent layers. The topology of GHSOM is presented in Fig. 2.

Here, we summarize the GHSOM algorithm:

(1) Initialize all parameters of GHSOM, including the learning rate, the neighborhood range, the initial map size for the training process, the growing-stopping criterion, the hierarchical stopping criterion, the maximum number of labels, and the label threshold.
(2) Start with a virtual layer 0 (see Fig. 2) consisting of only one unit whose weight vector is initialized as the average of all the input data. Then calculate the mean quantization error (mqe) by the Euclidean distance between the weight vector of the unit and all input vectors.
(3) Set the initial map size of the first layer in a small map of, for example, 2×2 units, which is self-organized according to the standard SOM training process.
(4) Evaluate the mapping quality by calculating the mean quantization error of each unit's mqe in the current layer to determine the *error unit* according to the largest deviation between its weight vector and the input vectors mapped to it. Then, either a new row or column of units is interpolated between the *error unit*

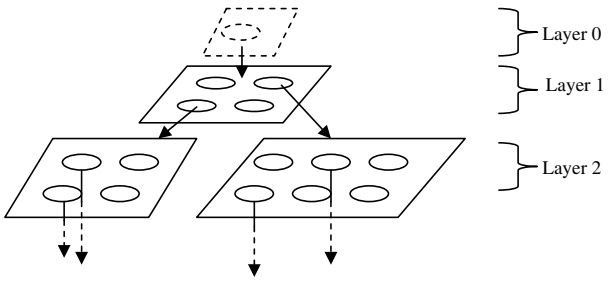

Fig. 2. Topology of GHSOM (Adapted from Dittenbach et al. (2002).

and its most dissimilar neighbor. The weight vectors of these new units are initialized as the average of their neighbors. Although the training process is very similar to the GSOM model, it uses a decreasing learning rate and a decreasing neighborhood range, instead of a fixed value. After growing the map, calculate the mean mqe of all units (MQE) in the current map. A map grows until its MQE is reduced to a predefined fraction (the growing-stopping criterion) of the mqe of the unit in the preceding layer of the hierarchy. In other words, the MQE of each map in the current layer should be smaller than a certain fraction value (τ_1) of the unit in the preceding layer. The lower the value of the quantization error, the better the map has been trained

$$\mathrm{MQE}_m < \tau_1 \cdot \mathrm{mqe}_u,$$

where m denotes the units in the current map and u denotes the mapped unit in the preceding layer.

(5) Determine the depth of each topic in the current layer according to a predefined fraction (τ_2) of the mqe of layer 0, i.e., the mqe of each unit in the current layer should be smaller than a certain fractional value of the unit in layer 0. The stopping criterion of any unit in the hierarchy is always compared with layer 0

$$\mathrm{mqe}_i < \tau_2 \cdot \mathrm{mqe}_0,$$

where i denotes the unit in the current layer.

The training procedure described in Steps 4 and 5 is used to train the subsequent layers.

3.3. The limitations of GHSOM models in representation

Although GHSOM overcomes the three drawbacks of SOM, it still lacks some important features. First, GHSOM may put some important topics in a deeper layer that is not obvious to users, so it may take a substantial amount of time to find such resources. Meanwhile, different users may give different weights to the importance of topics; thus, it is important to present the whole architecture in a brief way, such as using a tree display to indicate all the topics shown on maps. Another drawback of GHSOM is that we cannot determine the total volume of each topic (cluster) when we first view a map, but must browse through the map hierarchy to the bottom layer, i.e., the map cannot show the topic size of a legal map, which is especially important when there are a large number of documents. This could be overcome by showing the information about the number of topics on the map at the same time.

Although the GHSOM model performs the clustering task very well and the LabelSOM model can help by presenting topics with a list of labels, both lack the ability to clearly identify the topic name for each cluster automatically. Hence, in this paper, we propose a topic selection module that names a cluster automatically. This is discussed in detail in Section 4.

Hauck, Sewell, Ng, and Chen (2001) suggest that the use of color would make a map more visually interesting for users, but the GHSOM model does not use color displays. To overcome the above limitations of GHSOM, we propose an improved and integrated model in this paper.

4. Creating topic maps by using GHSOM

In this section, we describe each step of the system flow (Fig. 3).

4.1. Dataset – legal documents

We collected securities and futures related legal corpora from the Securities and Futures Institute (SFI), which is responsible for gathering, processing, and presenting securities and futures related information in Taiwan. SFI gathers most of its legal information from 14 organizations, such as SFB, TSE, OTC, Taifex, etc. These organizations are important participants in the stock and futures markets, because they play key roles in market management, promotion, surveillance, investor protection, etc. The dataset is composed of 959 legal documents (about 8000 web pages) promulgated by these organizations, and includes both current and abolished laws, rules, regulations, criteria, and contracts.

4.2. Term extraction

Unlike Western languages, word segmentation and term extraction are quite difficult in Chinese text processing. We extracted terms by implementing a maximum matching rule (Chen & Liu, 1992), based on a lexicon composed of a professional glossary provided by SFI and a common terms treasury. A total of 5571 terms were extracted from the documents.

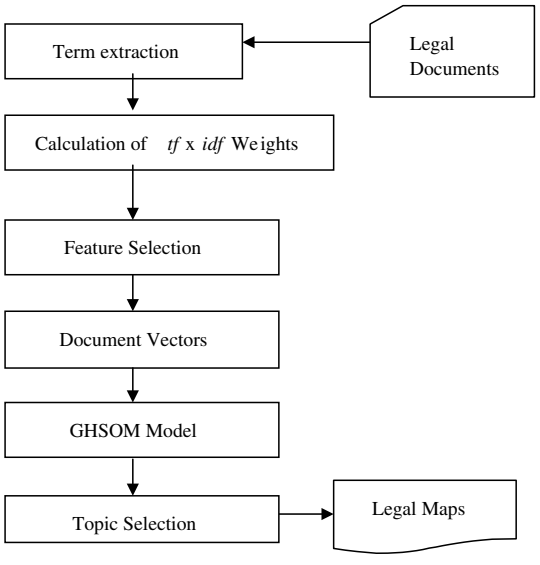

Fig. 3. System flow.

4.3. Calculation of tf × idf weights, feature selection and document vectors

We calculate the term frequencies of each document and the inverse document frequencies of each term. Then, a tf × idf weight vector (1), calculated by multiplying the term frequency by the inverse document frequency (Salton, 1989) and the word length of term for each document.[2]

$$w_i(d) = \mathrm{tf}_i(d) * \log\left(\frac{N}{\mathrm{df}_i}\right), \quad (1)$$

where $w_i(d)$ represents the weight of term (i) in document (d), $\mathrm{tf}_i(d)$ represents the number of times term (i) appears in document (d), N represents the total number of documents, df_i represents how many documents contain term (i).

The tf × idf weight vector is well-known in information retrieval and is considered the state of the art for content representation. This weighting scheme assigns high values to terms considered important for describing the contents of a document and discriminating between various documents. We select the features of the document vector on the basis of the top 2000 distinct term tf × idf values of all documents in our dataset. The input data of the neural network model, GHSOM, is represented by document vectors that can be described as the legal interpretation of a particular document.

4.4. GHSOM model

We use the input data generated by the above steps for GHSOM network training to produce a clustering architecture of the documents based on their contents. After a trial-and-error process, we set the parameters (defined in Section 3.2) as follows: $\tau_1 = 0.1$, $\tau_2 = 0.01$, and label threshold = 0.15. This completes the legal maps.

4.5. Topics selection

Because the label lists suggested by LabelSOM are not all suitable as topic names in this case, we designed a topic selection module to enhance the readability of topic names. We first obtained at most 20 labels for each cluster by the LabelSOM algorithm. The topic name for each cluster was selected from the top two highest-frequency labels based on the statistics of all article titles mapped to the cluster.

5. Browsing the legal maps

The training results are shown in Fig. 4, which illustrates the first layer of the legal maps generated by GHSOM. Each unit is assigned a set of up to five labels, based on the quantization error vector and the unit's weight vector.

Fig. 4. The top layer of legal maps.

The first layer of GHSOM seems to be clustered by functions of the markets, including the following 10 topics induced by the labels selected:

- Unit (1,1)[3]: Conformity with financial regulations.
- Unit (1,2): Protection of investors.
- Unit (1,3): Futures trading, clearing, and market surveillance.
- Unit (1,4): Futures trading assisted by securities firms.
- Unit (1,5): Information disclosure.
- Unit (2,1): Consignment of trading.
- Unit (2,2): Margin purchase and short sale of securities.
- Unit (2,3): Mutual fund related issues.
- Unit (2,4): Central custody of securities.
- Unit (2,5): Trading information transmission.

The geographical distance between Unit (1,1) and Unit (2,5) is the largest among all units, and is the same as the semantic distance between the two topics. Basically, conformity with financial regulations is unrelated to trading information transmission in the markets. The former relates to the financial transparency of companies that issue securities in securities markets, which is a market management concern. The latter belongs to trading information transmission, which focuses on the standards of hardware and software for trading information delivery.

Users can be further guided to browse more detailed topics in the first layer by clicking the "next level" hyperlink of each cluster. For example, in Unit (1,2) users can be guided to the second layer (see Fig. 5), which represents the following topics:

- Unit (1,1) and Unit (2,1): legal documents related to the securities and futures investor protection center, including mediation, business rules, compensation, assumpsit, and arbitration.

[2] Therefore, (1) can be rewritten as $w_i(d) = \mathrm{tf}_i(d) * \log\left(\frac{N}{\mathrm{df}_i} * I\right)$, where I represents the length of the term (i).

[3] We use the notation (x, y) to refer to the unit in row x and column y, starting with (1,1) in the upper left corner.

Fig. 5. The second layer of Unit (1,2) on the top legal map.

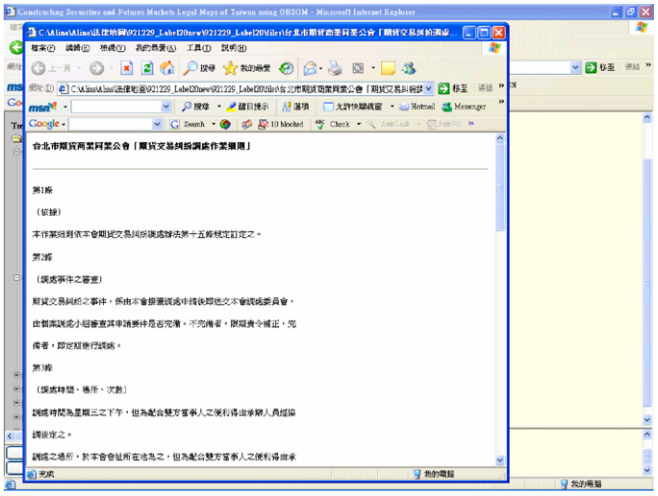

Fig. 7. The fourth layer of Unit (1,1) of the third legal map.

- Unit (1,2): dispute mediation.
- Unit (2,2): deliberation commission of IPO over TSE.
- Units (3,1) and (3,2): organization rules of institutes that promote both markets.

If users are interested in legal documents related to dispute mediation, they can be further guided to a deeper layer from Unit (1,2) in Fig. 5. By clicking on the "next layer" hyperlink of Unit (1,2) a user can see more detailed topics (see Fig. 6) that consist of all kinds of mediation rules promulgated by several related organizations whose purpose is to protect investors. In the last step, users can see the content of the legal document needed (Fig. 7) by clicking on the hyperlink in "Table of Contents" frame. For example, Unit (1,1) of the third layer contains the "Operating rules for futures trading dispute mediation" promulgated by the Taipei Futures Association. Both the content of the article (Fig. 7) and the features (labels) used to describe the article are presented to users. The labels selected as features of the "Operating rules for futures trading dispute mediation" are "party", "mediation", "Taipei Futures Association", and "dispute mediation". These labels seem to relate to this rule except for the term "party", because it often occurs in legal documents – especially in the contracts and documents related to disputes. This is clear from Fig. 6, where all the labels of the granularities contain the term "party".

6. Evaluation of the legal maps

We adopted two methods to verify the applicability of GHSOM to our dataset. First, we determined if legal documents regulating the same issue – but different aspects of it – could be clustered together automatically using a validity test. Second, we designed an experiment in which a task, to search some useful legal documents to find the proper channels (organizations) that could help testers settle a dispute between their futures brokers and themselves by using

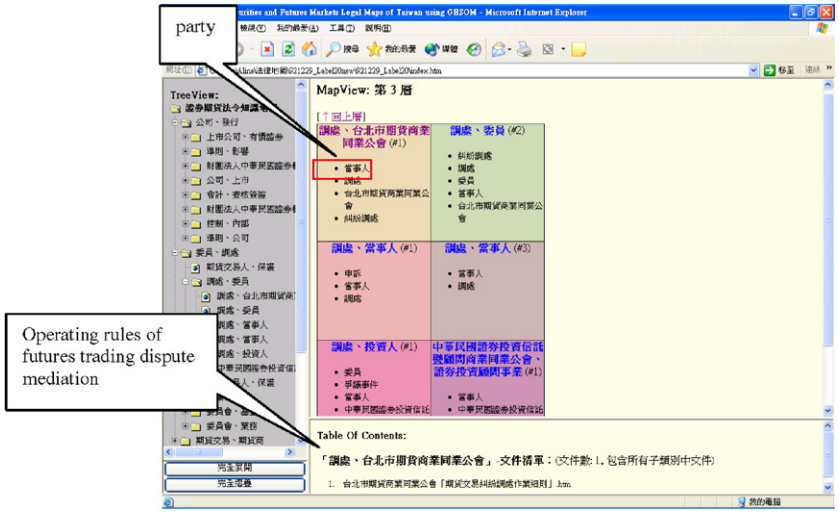

Fig. 6. The third layer of Unit (1,2) of the second legal map.

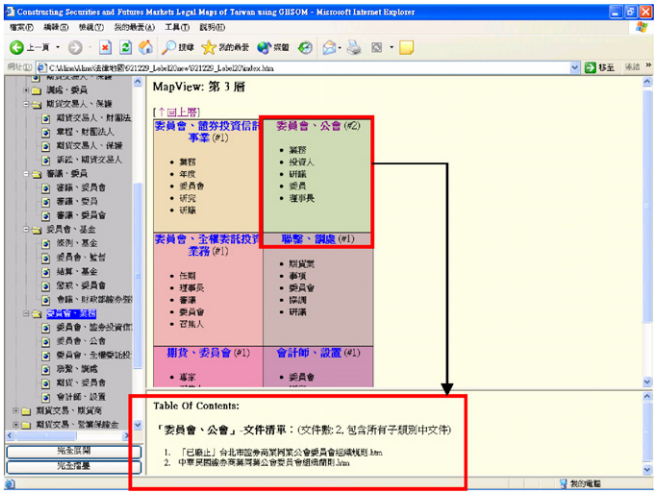

Fig. 8. Validity test sample-Unit (1,2) contains two similar sets of rules.

our system, was assigned to two groups of users to measure the value of the system.

6.1. Validity test

We demonstrate one of the validity tests with the following example. The content of Unit (1,2) in Fig. 8 includes two sets of organization rules, one for the Taipei Securities Association and the other for the Chinese Securities Association. The former was abolished and replaced by the latter in 1999. However, both sets of rules are very similar. In this example, GHSOM can cluster the two sets of rules in the same granularity.

6.2. A task experiment

We designed an experiment in which a task was assigned to a professional user and ten non-professional users to evaluate the value of the system. The professional user was a lawyer with expertise in financial markets. She was the benchmark with which we compared the non-professional users, who made up an experimental set. They (both the benchmark and the non-professional users) were asked to search some useful legal documents to find the proper channels (organizations) that could help them settle a dispute between their futures brokers and themselves. We required them to use our system to find the answer and record how long it took to accomplish the task. The results show that the time spent by the benchmark was 2.5 min and the average time spent by non-professional users was 5 min. In other words, using our system the non-professional users only needed twice as much time as the lawyer to accomplish the task.

7. Conclusions

Legal knowledge representation of the securities and futures markets can help market participants understand the relationships between the laws, regulations, rules, criteria, and contracts in both markets. Access to legal document archives would be enhanced by providing investors with effective tools to search an environment of dynamically classified legal documents. In this paper, we integrate a dictionary-based term extraction technology, a vector space model for processing documents, and the GHSOM and LabelSOM algorithms to construct legal maps. Our method enables users to access legal documents in a convenient, efficient, and cost-effective way. We also enhance the visualization and readability of the GHSOM model by adding a tree display to show the topics generated by our topic selection module. In addition, information regarding the volume of documents clustered in each topic and a color display are shown on the maps to improve visualization. The maps can be browsed on the web site (http://synteny.iis.sinica.edu.tw/legalmap/).

In this study, clustering of legal documents using the GHSOM model is shown to be effective, because most similar documents can be clustered together and the geographic distance between clusters seems to also represent the semantic distance. The results show that the model performs well in clustering Chinese language legal documents.

In our experience, term extraction is the only critical difference between different language legal informatics. It is hard to identify terms in Chinese articles, because there are no spaces between terms. Some methodologies have been developed to resolve the problem, such as dictionary, linguistic, and statistical approaches (Ong et al., 1999).

With regard to the clustering method, GHSOM, we find that there is no difference between Chinese and Western language formats, because the input vectors are composed of $tf \times idf$ weights generated by calculating the term and document frequencies. Also, the terms used have already been extracted by a term extraction technology.

To evaluate the applicability of GHSOM to legal maps, we invited 15 experts[4] to use the system and give some suggestions to us. After reviewing the respondents opinions, some users remarked that the distinction between nearby clusters is not clear enough for them to browse legal maps and needs to be improved, therefore, we suggest that future research should pay more attention to clarifying some terms that appear in most legal documents (e.g., "party" and "securities") in order to make the distinction between nearby clusters on the legal maps more obvious to users. This would reduce the difficulties that many users experience when browsing maps. Besides, they also suggested that the addition of keyword search and notification about amendments to laws, as well as more information about banking and insurance related laws and regulations, would

[4] We invited 15 people who work in Taiwan's securities and futures markets, to answer the questionnaire. The group consisted of lawyers, government staff, others associated with securities firms and banks, and individual investors.

make the system easier to retrieve legal information about the securities and futures markets.

In legal informatics of oriental languages, we suggest that future research could apply other major methodologies from legal informatics of Western languages by developing an ontology defined in terms of specific users. For example, an ontology could be developed for people working in financial holding companies, and used to constructing a knowledge base for their specific purposes, including consumer financing, corporate financing, wealth management, etc.

References

Anandanpillai, T., & Barta, T. A. (1999). A case-based reasoning system for housing discrimination law. *Expert Systems with Applications, 16*, 315–324.

Chen, K. J., & Liu, S. H. (1992). Word identification for mandarin Chinese sentences. In *Proceedings of the fifteenth international conference on computational linguistics nantes (COLING92)*.

Daniels, J., & Rissland, E. (1995). Finding legally relevant passages in case opinions. In *Proc international conference on artificial intelligence and law (ICAIL)* (pp. 39–46).

Dittenbach, M., Rauber, A., & Merkl, D. (2002). Uncovering hierarchical structure in data using the growing hierarchical self-organizing map. *Neurocomputing, 48*, 199–216.

Elhadi, M. T. (2000). Bankruptcy support system: taking advantage of information retrieval and case-based reasoning. *Expert Systems with Applications, 18*, 215–219.

Fritzke, B. (1995). Growing grid – a self-organizing network with constant neighborhood range and adaptation strength. *Neural Processing Letters, 2*(5), 9–13.

Hauck, R. V., Sewell, R. R., Ng, T. D., & Chen, H. (2001). Concept-based searching and browsing: a geoscience experiment. *Journal of Information Science, 27*(4), 199–210.

Honkela, T., Kaski, S., Lagus, K., & Kohonen, T. (1997). WEBSOM-self-organizing maps of document collections. In *Proc. workshop on self-organizing maps (WSOM97), Espoo, Finland.*

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps, Biol. *Cybernet, 43*, 59–69.

Kohonen, T. (1995). *Self-Organizing Maps*. Berlin: Springer.

Merkl, D., & Schweighofer, E. (1997). *The exploration of legal text corpora with hierarchical neural networks: a guided tour in public international law (ICAIL)*. ACM Press.

Merkl, D., Schweighofer, E., & Winiwarter, W. (1995). Analysis of legal thesauri based on self-organizing feature maps. In *Proc. fourth international conference on artificial neural networks (ANN'95), Cambridge*.

Moens, M. (1999). *Automatic indexing and abstracting of document texts*. Boston: Kluwer Academic Publishers.

Ong, T., & Chen, H. (1999). Updateable Pat-tree approach to Chinese key phrase extraction using mutual information: a linguistic foundation for knowledge management. In *Proceedings of the second Asian digital library conference, Taipei, Taiwan*.

Pal, K. (1999). An approach to legal reasoning based on a hybrid decision-support system. *Expert Systems with Applications, 16*, 1–12.

Rauber, A. (1999). LabelSOM: on the labeling of self-organizing maps. In *Proceedings of the international joint conference on neural networks (IJCNN'99), Washington, DC*.

Rodrigues, J. S., & Almeida, L. B. (1990). Improving the learning speed in topological maps of patterns. *Proceedings of INNC*, 813–816.

Salton, G. (1989). *Automatic text processing*. MA: Addison-Wesley.

Schweighofer, E., Rauber, A., & Dittenbach, M. (2001). *Automatic text representation classification, and labeling in European law, ICAIL*. ACM Press.

Schweighofer, E., & Winiwarter, W. (1995). KONTERM: exploratory data analysis for semi-automatic indexation of legal documents. In N. Revell, & A.M. Tjoa (Eds.), *Proc. 6th database and expert systems applications DEXA'95 workshop* (pp. 407–412).

Schweighofer, E., Winiwarter, W., & Merkl, D. (1995). Exploratory data analysis as a mean of structuring legal knowledge about concepts and documents. In: Pre-Proc. 17th IVR world congress 1995, Bologna, challenges to law at the end of the 20th century. *European Journal of Law, Philosophy and Computer Science*, Bologna (pp. 400–409).

Smith, J., Maccrimmon, D., Atherton, B., McClean, J., Shinehoft, J., & Quintana, L. (1995). Artificial intelligence and legal discourse: the Flexlaw legal text management system. *AI and Law, 3*(1), 55–95.

Thompson, P. (2001). *Automatic categorization of case law, ICAIL*. ACM Press.