Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 2290–2297
Expert Systems with Applications
www.elsevier.com/locate/eswa

Structure clustering for Chinese patent documents

Su-Hsien Huang [a,d,*], Hao-Ren Ke [b], Wei-Pang Yang [c]

[a] Institute of Computer Science and Engineering, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan, ROC
[b] Library and Institute of Information Management, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan, ROC
[c] Department of Information Management, National Don Hwa University, 1, Section 2, Da Hsueh Road, Shou-Feng, Hualien, Taiwan, ROC
[d] Department of Information Management, Minghsin University of Science and Technology, 1, Hsin Hsin Road, Hsin Feng, Hsinchu, Taiwan, ROC

Abstract

This paper aims to cluster Chinese patent documents with the structures. Both the explicit and implicit structures are analyzed to represent by the proposed structure expression. Accordingly, an unsupervised clustering algorithm called structured self-organizing map (SOM) is adopted to cluster Chinese patent documents with both similar content and structure. Structured SOM clusters the similar content of each sub-part structure, and then propagates the similarity to upper level ones. Experimental result showed the maps size and number of patents are proportional to the computing time, which implies the width and depth of structure affects the performance of structured SOM. Structured clustering of patents is helpful in many applications. In the lawsuit of copyright, companies are easy to find claim conflict in the existent patents to contradict the accusation. Moreover, decision-maker of a company can be advised to avoid hot-spot aspects of patents, which can save a lot of R&D effort.

Keywords: Structure clustering; Chinese patent; Structure expression; Metadata

1. Introduction

1.1. Background

Digital libraries provide various integrated services to coordinate information over Internet. However, the distributed and heterogeneous data complicate the integration when it involves several issues, such as format, content, semantics, etc. The discrepancy and redundancy confuses users in readability. To provide a unified view, data clustering with an overview by integrating similar data has received considerable attention in recent researches.

Conventional data clustering adopts feature vector to represent data. It lacks some aspects of consideration to identify similarity. For example, chemical compounds with the same molecules but different structures are not the same. Moreover, two trademarks with the same components but different placements cannot be identified similar, too. These two examples account for the structure is a key-factor to influence the clustering in particular domains. Although conventional structure clustering focuses on visual pattern query, chemical structure identification and syntactic parsing tree, rare mentioned in documents. This is owing to the structure in documents is hard to obtain. Interestingly, the development in rhetorical structure theory (RST) facilitates the extraction of structure. RST was proposed in the 1980s and has been successfully applied to documents summarization, automatic layout, and so on. RST categorizes the characteristics of phrases and constructs a closure tree to represent structure. Therefore, applying RST to represent documents structure becomes possible and practical.

Several kinds of documents, like patent and electronic thesis, contain both structure and content information. To cluster these documents, only the same content in the same structure can be identified similar. The tort of copyright is a good illustration to require conflict in both claim

* Corresponding author.
E-mail address: sshuang@cis.nctu.edu.tw (S.-H. Huang).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.03.012

and claim structure. The structure of documents can be categorized into two types. Explicit structure represents the existent attributes of documents, like "subject", "abstract", "publisher" and "classification" in patent documents. On the other hands, the structure analyzed from content is implicit structure. The implicit structure can be extracted by analyzing the writing style and the well-established convention. For example, in the claim field of patents, "the ... of claim 1", "comprising", "wherein", "having", "consist of", etc. construct the implicit structure. With both explicit and implicit structures, document structure can be constructed and applied in structure clustering.

This study tries to apply Chinese patent documents in structure clustering. Patent structure has been mentioned in recent literatures. Shinmori, Okumura, Marukawa, and Iwayama (2003) provides a rhetorical method to categorize Japanese patent structure into three styles – process sequence style, element enumeration style and Jepson-like style. Each claim in the patent is given a weight to evaluate the relationships among different claims. Fujii, Iwayama, and Kando (2006) extends Shinmori's work and adds delimiter to punctuate claim into components. Mase, Matsubayashi, Ogawa, Iwayama, and Oshio (2004) analyze the patent structure into premise, description and target parts. Keywords in different part can be re-weighted to enhance query precision (Iwayama, Fujii, Kando, & Marukawa, 2006).

Clustering in patent is a real-time application. Real-time means the time constraint of clustering algorithm is tight when users require the result on-line. Unfortunately, structure clustering is more complicate than conventional one, where the computing time is subject to the structure. Hence, the first requirement of structure clustering is efficiency. Additionally, real-time also implies there is no training data in advance. Therefore, unsupervised algorithm is suggested when there is unnecessary to use training data in the process. Unsupervised clustering integrates similar data in definite circles, and adjusts the parameters to obtain result. Kohonen proposes an unsupervised neural network self-organizing map (SOM) and receives excellent performance (Kohonen, 1998). SOM provides unsupervised neural network clustering and maps high dimension data (usually two) into a low-dimension map. In SOM, closer nodes in the map imply shorter distance in real data. SOM applies in many domains, like bio-structure clustering, graph structure clustering and audio-pattern clustering, etc. The multi-dimension representation of SOM facilitates the readability which is suitable for patent clustering.

1.2. Literatures of early studies

It is noteworthy that conventional SOM cannot deal with structure data. In 1998, a general framework is presented to deal with structure by neural network (Frasconi, Gori, & Sperduti, 1998). This framework constructs directed ordered acyclic graphs (DOAG) for structure data and recurrently proceeds data by following DOAG sequence. This framework further motivates the upcoming researches. SOMSD (Sperduti, 2001) applies self-organizing map (SOM) to manipulate structured data. Each neuron in SOMSD is equipped with two weights: one for previous structure and one for two-time previous structure. Structure similarity is estimated by concatenating current similarity and these two structure similarity. RSOM (Voegtlin, 2002) recursively calculates SOM by adding current similarity and previous structure. Hammer, Micheli, and Sperduti (2002) propose a general framework to further extend SOMSD. Vesanto and Alhoniemi (2000) cluster each sub-structure part in single SOM and use hierarchical k-means cluster to investigate similar structure. Smith and Ng (2003) and Roussinov and Chen (2001) derive user navigation patterns as structure to cluster web page navigation pattern. Similar navigation patterns (treated as structure graph) are clustered together by SOM. Xu, Chang, and Paplinski (2005) use on-line expansion to enlarge SOM size into multiple layers. Other of structure clustering can be found in literatures (Hagenbuchner, Sperduti, & Tsoi, 2004; Hammer & Jain, 2004; Rossi, Conan-Guez, & Golli, 2004; Sperduti & Starita, 1997; Strickert & Hammer, 2003).

1.3. Purpose of this study

This study is intended to apply Chinese patent documents in structure clustering. A patent structure is analyzed first and provides structure expression to represent it. Having the structure established, self-organizing map (SOM) is adopted. Several modification of SOM is undertaken to process structure, including input, output and training algorithm. In brief, three major objectives involved in this study are:

(1) Construct expression model to represent the patent structure.
(2) Develop structure clustering algorithm to process structure patent.
(3) Evaluate the structure clustering algorithm.

This paper is organized as follows. Chapter 2 analyzes the structure of Chinese patent documents. Chapter 3 proposes structured SOM to cluster patents. Chapter 4 implements structured SOM and a series of experiments are conducted in Chapter 5. Chapter 6 draws the conclusions.

2. Structure analysis

As shown in Fig. 2, the preprocessing of structure clustering is to analyze the structure. The received documents are first represented by structure expression, and then determine the input sequence and maximum branch number (MBN). To formally describe the structure, two formal constructors are defined. Given a *Structure S*, S can be represented by the following two constructors:

(1) *Tuple constructor* (*TC*): A tuple constructor TC of Structure S is an ordered list constructed by the union of single-value attributes. For example, the set of attributes c_i, $c_1 \cdot \text{occurence} = 1$, $c_2 \cdot \text{occurence} = 1$, $\ldots, c_n \cdot \text{occurence} = 1$, are represented as $\text{TC}_\tau = \{c_1, c_2, \ldots, c_n\}_\tau$. The subscript τ is the name of TC.
(2) *Set constructor* (*SC*): A set constructor SC of Structure S is a multi-value type with the same occurrence larger than one. For example, the set of attributes c_i, $c_1 \cdot \text{occurence} = i$, $c_2 \cdot \text{occurence} = i, \ldots, c_n \cdot \text{occurence} = i$, are represented as $\langle c_1, c_2, \ldots, c_n \rangle_\tau^i$. The subscript τ represents the name of SC and i represents the maximum occurrence.

2.1. Explicit structure

A structure document contains two types of structure. The first one is *explicit structure*, which is obtained from the schema of structure documents (like data metadata). Explicit structure is applicable in documents with regular schema. To take a simple example of electronic thesis, the attributes "subject", "abstract", "chapter", "section", and "paragraph" contain explicit structure. The structure expressions above example can be expressed as follows:

$$\left\{ \text{Subject}, \text{Abstract}, \left\langle \left\langle \langle \text{Content} \rangle_{\text{paragraph}}^{n1} \right\rangle_{\text{section}}^{n2} \right\rangle_{\text{chapter}}^{n3} \right\}_{\text{Book}}$$

where $n1$, $n2$ and $n3$ stand for the maximum occurrence of the attributes.

Applying explicit structure in structure clustering is meaningful when the documents contain regular schema and massive content in each attribute. For example, electronic thesis and XML-based news are suitable for explicit structure clustering. In this study, Chinese patent documents contain most information in "Claim" field. Therefore, implicit structure is introduced in the next section.

2.2. Implicit structure

The structure analyzed from content is *implicit structure*. The implicit structure can be extracted by analyzing the writing style and the well-established convention. In the "Claim" field of Chinese patent documents, two types are categorized:

- *Composition style*: As in "包括" (comprising), "包含" (includes), the set of element is described. These keywords are used in method to imply the composition of the inventions. This type of style represents a set of elements contains in the claim structure.
- *Pre-condition style*: As in "如….所述" (as claimed in …), "其中…" (wherein), these descriptions imply the statements has followed a list of composition. This type of style represents a list of compositions construct the claim structure.

Comparing with other-linguistic patents, there are substantial difficulties to obtain several relationships in Chinese patent. For example, in Shinmori's method, the *process sequence style* (means the sequence of processes, like "does", "and does") is ambiguous in Chinese grammar. It's always followed by "一個…" where is the same meaning with the article "one".

These two types of style can derive the implicit structure of patents. For the presence of these keywords, a hierarchy relationship can be constructed. For example, Fig. 1 illustrates an example of Chinese patent and represents the implicit structure analyzed by these two styles. To give each

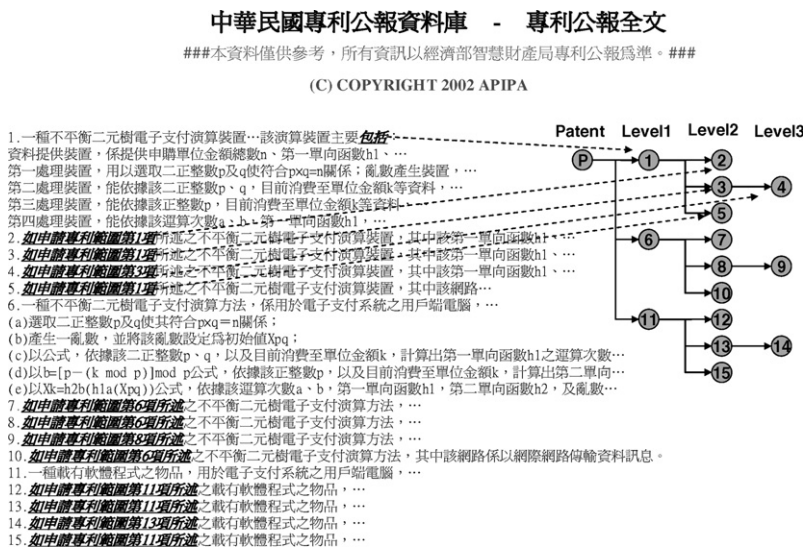

Fig. 1. Structure of Chinese patent.

segmented claim an ID, naming mechanism is defined for each claim as follows:

$$\text{Patent\_ID-Claim\_ID}_{\text{level1}}\text{-}\cdots\text{Claim\_ID}_{\text{level}i}$$

The structure can be formally described by the structure expression, which can be more comprehensive for structure clustering. The composition style means the occurrence happens in the following statements. By following the structure expression in previous section, it can be expressed as $\langle\rangle_i$, where i stands for the structure level. The pre-condition style implies an attribute contains in the following statement. By following the structure expression, it can be expressed as $\{\}_i$, where i stands for the structure level. For the example of Fig. 1, the structure expression can be expressed as follows:

$$\left\langle \{\text{Claim}_a, \{\text{Claim}_b\}_{\text{level2}}, \text{Claim}_c\}_{\text{level1}} \right\rangle^3_{\text{patent}}$$

2.3. Determine input sequence

After the structure is analyzed, the input of each sub-part structure is determined by following *directed acyclic graph* (*DAG*). A depth-first search is adapted in the navigation of structure tree.

For the formal expression of both explicit and implicit structure, the input of structure follows two principles:

(1) By order of the attributes in tuple constructor.
(2) Iteratively expand set constructor to i times.

For example, the input sequence of explicit structure

$$\left\{ \text{Subject}, \text{Abstract}, \left\langle \left\langle \langle \text{Content} \rangle^{n1}_{\text{paragraph}} \right\rangle^{n2}_{\text{section}} \right\rangle^{n3}_{\text{chapter}} \right\}_{\text{Book}}$$

is

$$\begin{aligned}\text{Subject} \rightarrow \text{Abstract} \rightarrow \text{Chapter}^1 \rightarrow \text{Section}^1 \\ \rightarrow \text{Paragraph}^1 \rightarrow \cdots \rightarrow \text{Paragraph}^{n3} \rightarrow \text{Section}^2 \\ \rightarrow \cdots \rightarrow \text{Paragraph}^{n2} \rightarrow \cdots \rightarrow \text{Chapter}^{n1} \\ \rightarrow \text{Section}^{n2} \rightarrow \text{Paragraph}^{n1}\end{aligned}$$

For the example of Fig. 1, the input sequence to structure clustering is:

$$P \rightarrow 1 \rightarrow 6 \rightarrow 11 \rightarrow 2 \rightarrow 3 \rightarrow 5 \rightarrow 4 \rightarrow 7 \rightarrow 8 \rightarrow 10 \rightarrow 9 \rightarrow 12 \rightarrow 13 \rightarrow 15 \rightarrow 14$$

2.4. Determine maximum branch number

The next step is to determine the *maximum branch number* (*MBN*) of *tuple and set constructor*. The maximum branch number (MBN) represents the maximum number of branches in patent structure. In the structure expression of both explicit and implicit, MBN is the maximum number of TC attributes and SC occurrence. For example in Fig. 1, the MBN is 3 when the largest branch and occurrence is less than 3.

3. Structured SOM

Self-organizing map (*SOM*) (Kohonen, 1998) provides unsupervised neural network clustering and maps high-dimension data into a low-dimension map (usually two). There are five steps to train SOM (in Fig. 2):

(1) Initialize weight vectors of output map as the same number features with input document vector.
(2) Present input documents in order.

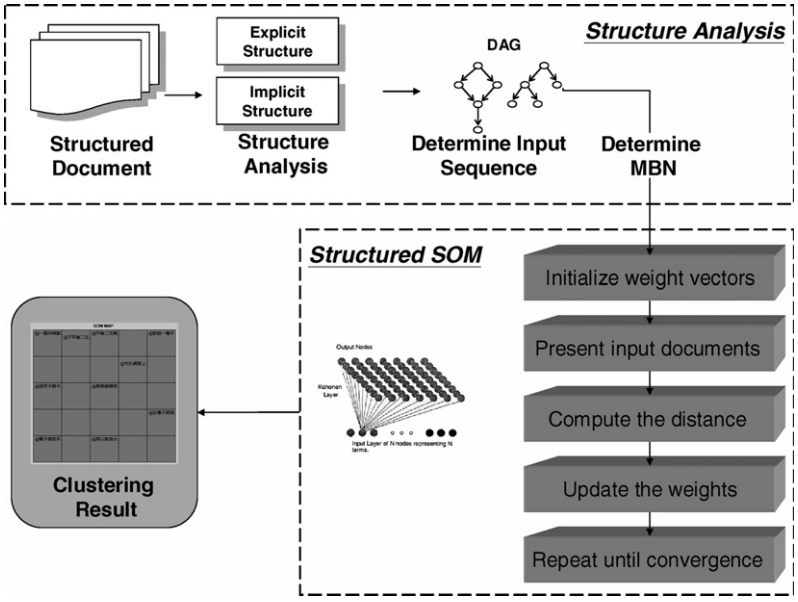

Fig. 2. Structured SOM.

(3) Compute the distance between the input document and all nodes in the map and select the closest node as the winner.
(4) Update the weights of the winner node and its neighbors.
(5) Repeat steps (3)–(4) to other documents and iterate all inputs until convergence. Label the regions of the final map to represent the clustering result.

SOM with structure receives considerable attention in recent years (Rossi et al., 2004; Strickert & Hammer, 2003; Voegtlin, 2002). The structured SOM in this study refers to Hagenbuchner's self-organizing map clustering algorithm in structured data (Hagenbuchner et al., 2004). Hagenbuchner's approach applied structured SOM on the images identification. The main concept is to calculate each sub-structure respectively, and concatenate the result to the upper structure. Notably, the images in Hagenbuchner's method have simple structure and regular components. Experiment result has shown that structured SOM can cluster similar images together and distinguish the structure difference in the map. This paper applies structured SOM in Chinese patents and tries to identify similar patent with different structure. The process of structured SOM is described in Fig. 2.

3.1. Training for explicit structure

Applying SOM in Chinese patent documents requires modification of input/output vectors and algorithm. The explicit structure can be represented by structure expression shown in previous section. The next step is to determine the maximum branch number (MBN) of tuple and set constructor in explicit structure. The shortage of nodes less than MBN requires additional *Null* nodes. For the simplicity, assume the MBN = 3 in the example of explicit structure ($n1 = 3$, $n2 = 3$, $n1 = 1$), the structure expression can be rewrite as:

$$\left\{\text{Subject},\text{Abstract},\left\langle\left\langle\langle\text{Content},\text{NULL},\text{NULL}\rangle^{n1}_{\text{paragraph}}\right\rangle^{n2}_{\text{section}},\right.\right.$$
$$\left.\left.NULL,NULL\rangle^{n1}_{\text{chapter}}\right\}_{\text{Book}}$$

The output nodes of SOM also need to be modified as the same number fields as input vector. For example, the output node of structured SOM in previous example is

$$d_{x,y} = (V_{d_{x,y}}, (x_1, y_1), (x_1, y_1), (x_1, y_1))$$

The structure expression is assigned into SOM input vector by directed acyclic graph (DAG) sequence, for example, given a three node input vector, the input vector are rewrite as following:

$$d_{\text{book}} = (V_{\text{book}}, (x_{\text{subject}}, y_{\text{subject}}), (x_{\text{abstract}}, y_{\text{abstract}}), (x_{\text{chapter}}, y_{\text{chapter}}))$$

Additionally, the distance calculation is updated as following:

$$d = \sqrt{(V_{\text{Node0}} - V_{\text{book}})^2} + |V_{\text{Node1}} - V_{\text{subject}}| + |V_{\text{Node2}} - V_{\text{abstract}}| + |V_{\text{Node3}} - V_{\text{chapter}}|$$

where $|V_{\text{Node}i} - V_{\text{Node}j}|$ represents the distance between these two vectors. The modification of distance formula means the cascaded calculation to all connected nodes. The adapting of structured SOM is updated as following:

$$w_{d_{x,y}}(t+1) = w_{d_{x,y}}(t) + \eta(t) \times |V_{\text{Node1}} - V_{\text{subject}}|$$
$$w_{d_{x_1,y_1}}(t+1) = w_{d_{x,y}}(t) + \eta(t) \times |V_{\text{Node2}} - V_{\text{abstract}}|$$
$$w_{d_{x_2,y_2}}(t+1) = w_{d_{x,y}}(t) + \eta(t) \times |V_{\text{Node3}} - V_{\text{chapter}}|$$

3.2. Training for implicit structure

For the implicit structure, the dimension of the vector is also determined by the MBN in all input patents. Assume the MBN = 3, the representation in previous example is:

$$(V_{\text{Patent}}, \text{Claim}_1, \text{Claim}_6, \text{Claim}_{11})$$
$$(V_{\text{Node1}}, \text{Claim}_2, \text{Claim}_3, \text{Claim}_5)$$
$$(V_{\text{Node6}}, \text{Claim}_7, \text{Claim}_8, \text{Claim}_{10})$$
$$(V_{\text{Node11}}, \text{Claim}_{12}, \text{Claim}_{13}, \text{Claim}_{15})$$

The output nodes of SOM also need to be modified as the same number fields as input vector. For example, the output node of structured SOM in previous example is

$$d_{x,y} = (V_{d_{x,y}}, (x_1, y_1), (x_1, y_1), (x_1, y_1))$$

Additionally, the distance calculation is updated as following formula:

$$d = \sqrt{(V_{\text{Node0}} - V_{\text{Patent}})^2} + |V_{\text{Node1}} - V_{\text{Claim1}}| + |V_{\text{Node2}} - V_{\text{Claim6}}| + |V_{\text{Node3}} - V_{\text{Claim11}}|$$

where $|V_{\text{Node}i} - V_{\text{Node}j}|$ represents the distance between these two vectors. The modification of distance formula means the cascaded calculation to all connected nodes. The adapting of structured SOM is also updated as following:

$$w_{d_{x,y}}(t+1) = w_{d_{x,y}}(t) + \eta(t) \times |V_{\text{Node1}} - V_{\text{Claim1}}|$$
$$w_{d_{x_1,y_1}}(t+1) = w_{d_{x,y}}(t) + \eta(t) \times |V_{\text{Node2}} - V_{\text{Claim6}}|$$
$$w_{d_{x_2,y_2}}(t+1) = w_{d_{x,y}}(t) + \eta(t) \times |V_{\text{Node3}} - V_{\text{Claim11}}|$$

4. Implementation

Fig. 3 displays the structured SOM system. The system contains six major parts:

(1) *SOM system*: Choose standard SOM and structured SOM.
(2) *Parameter setting*: Set the map dimension, learning rate, adaptation neighborhood and iteration to train.
(3) *Function selection*: Select open file, close result, training, save result, resize map and quit training.

Fig. 3. Implementation of structured SOM.

(4) *Map information*: Display the containing patents in the SOM grid.
(5) *System message*: Show the current iteration and processing message.
(6) *SOM map*: Illustrate the clustering result.

Fig. 4 demonstrates an example to apply structured SOM in Chinese patent documents. Initially, 10 Chinese patent documents are selected from database. These documents are sent to standard SOM to obtain preliminary clustering. In the clustering, totally four clusters are labeled, *e-commerce, binary-tree, identification* and *database*. The interesting is paid in the five patents of *e-commerce* cluster. There are three observations:

(1) Commerce system is subjected.
(2) Client–server architecture is addressed.
(3) Encrypt system is applied.

Hereafter, these patents are sent to structured SOM. In the structured SOM, these five patents are separated into three clusters, which represent they are different in structure. For *cluster 1*, the patent primarily describes the con-

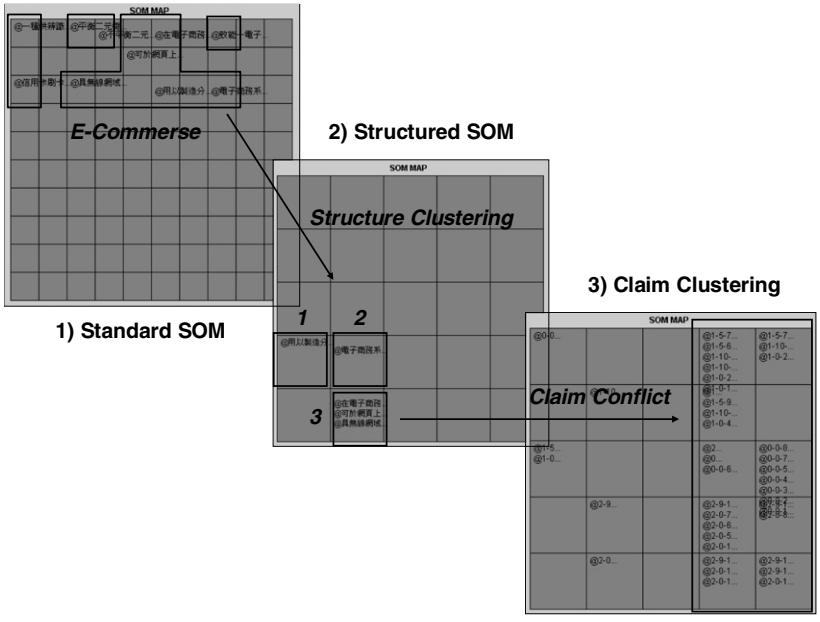

Fig. 4. Patent clustering.

struction of commerce system and is written in a very flat style. On the contrary, *cluster 2* mainly focuses on encrypt system and is written in a very deep style. These two clusters are distinguished with cluster 3 not only the different subjects but the different patent structure. Notably, three documents marked as *cluster 3* clusters together and might have high risk to conflict in the claims. Therefore, these claims are segmented by following the naming mechanism and cluster by SOM again to obtain conflict clusters. In the third step of Fig. 4, claim conflicts have observed in three areas:

(1) Wire and wireless.
(2) Client-side and server-side service.
(3) Digital authorization and public key encrypt.

These three patent conflicts provide good advisement for patent inventors to avoid involvement in this area. Structured SOM is also helpful in many patent applications. In the lawsuit of copyright, companies are easy to find claim conflict in the existent patents, which is easy to contradict the accusation. Moreover, decision-maker of a company can be advised to avoid hot-spot aspects of patents from structured SOM, which can save a lot of R&D effort.

5. Experiments

A series of experiments are conducted to examine the performance of structured SOM. The data set comes from the claim attribute of Chinese patent documents, to examine the implicit structure of patents. The experimental platform is Intel Celeron 1.5 GHz CPU, 512 MB RAM, Microsoft 2000 OS. The structured SOM was implemented by Java.

In Table 1, the experiment was conducted to examine the performance in different map size. The parameters were set to 10 documents in 50 iterations with MBN = 9. The computing time of both standard and structured SOM is positive proportional to the map size. However, structured SOM is more time-consuming than standard one.

In Table 2, the experiment was conducted to examine the performance in different document numbers. The parameters were set to 5×5 map in 50 iterations with MBN = 9. The computing times of both standard and structured SOM are positive proportional to the document numbers. The experiment corresponding to Table 1 implies the map size and documents are important factors to influence the computing time. However, the time complexity between standard and structured SOM is extremely large, too. One explanation for the phenomenon is the document structure complicates the clustering. The deeper the structure is, the longer the computing time.

Table 1
Time vs. map size (s)

Map size	2×2	3×3	5×5	10×10
Standard SOM	1	6	29	114
Structured SOM	227	355	2927	11,932

(10 documents, 50 iterations, MBN = 9).

Table 2
Time vs. documents (s)

Document number	5	10	20
Standard SOM	8	29	118
Structured SOM	1727	2927	5713

(5×5 map, 50 iterations, MBN = 9).

Table 3
Times vs. MBN (s)

MBN	3	5	6	9
Structured SOM	7	10	53	295

(5×5 map, 10 documents, 5 iterations).

The experimental result of Table 3 further introduced another factor to influence computing time. With 10 documents in five iteration of 5×5 map, structured SOM is nearly exponential proportional to the MBN. This is the most important factor account for the complexity of structured SOM because both input vector and output map are modified to fit the dimension of MBN. These results lead to the conclusion that the width (MBN) and the depth of the structure are two key-factors for structured SOM performance.

6. Conclusions and future work

Increasing Chinese patent documents require the clustering in the application. To cluster precisely, patents with similar content but different structure should be distinguished. However, conventional clustering lacks of structure consideration cannot identify similar patents with different structure, and rare research mentioned in text domain. In this paper, a clustering algorithm called structured SOM is proposed to clusters structure patents by considering the content and structure information simultaneously. Structured SOM requires the analysis of patent structure in advance. Two types of structure expression, explicit and implicit structure, are provided. Structured SOM modifies the input and output vectors for structure documents and iteratively training by adapting each subpart of structure.

An example to cluster Chinese patent documents has successfully implemented. Claim conflict appeared in the analysis provides advisements for patent inventors and decision-makers of business. In the lawsuit of copyright, companies are easy to find claim conflict in the existent patents, which is easy to contradict the accusation. Moreover, decision-maker of a company can be advised to avoid hot-spot aspects of patents from structured SOM, which can save a lot of R&D effort.

Experiments are conducted to examine the performance of structured SOM. Conclusions are given that both map

size and documents are positive proportional to the computing time. This implies two factors are important: the width (MBN) and depth of structure. MBN is nearly exponential drop-off the performance of structured SOM. In some practical application, prediction is adapted to estimate node distance to reduce computing time but sacrifice neglectful accuracy.

Structure clustering is rarely mentioned in the text domain because the text with structure (like data metadata) has small amount of content in the attributes (explicit structure). With the development of digitalization, some structure documents, like electronic thesis and on-line news, can be applied in structured SOM. The future direction for this study might be to discover more applications for structure clustering. It is also lots of work to improve the efficiency of structured SOM.

Acknowledgements

This research was supported by the Software Technology for Advanced Network Application project of Institute for Information Industry in 2004 and sponsored by MOEA, ROC.

References

Frasconi, P., Gori, M., & Sperduti, A. (1998). A general framework for adaptive processing of data structures. *IEEE Transactions on Neural Networks, 9*(5), 768–786.

Fujii, A., Iwayama, A., & Kando, N. (2006). Test collections for patent retrieval and patent classification in the fifth NTCIR workshop. In *Proceedings of the fifth international conference on language resources and evaluation* (pp. 671–674).

Hagenbuchner, M., Sperduti, A., & Tsoi, A. C. (2004). A self-organizing map for adaptive processing of structured data. *IEEE Transactions on Neural Networks, 14*(3), 491–505.

Hammer, B., & Jain, B. J. (2004). Neural methods for non-standard data. In *european symposium at artificial neural networks 2004* (pp. 281–292).

Hammer, B., Micheli, A., & Sperduti, A. (2002). A general framework for unsupervised processing of structured data. In *European symposium on artificial neural networks (ESANN'2002)* (pp. 395–400).

Iwayama, M., Fujii, A., Kando, N., & Marukawa, Y. (2006). Evaluating patent retrieval in the third NTCIR workshop. *Information Processing and Management, 42*(1), 207–221.

Kohonen, T. (1998). The self-organizing map. *Neurocomputing, 21*, 1–6.

Mase, H., Matsubayashi, T., Ogawa, Y., Iwayama, M., & Oshio, T. (2004). Two-stage patent retrieval method considering claim structure. In *NTCIR workshop 4 meeting*.

Rossi, F., Conan-Guez, B., & Golli, A. E. (2004). Clustering functional data with the SOM algorithm. In *Proceedings of european symposium on artificial neural networks 2004 (ESANN'04)* (pp. 305–312).

Roussinov, D. G., & Chen, H. (2001). Information navigation on the web clustering and summarizing query results. *Information Processing and Management, 37*, 789–816.

Shinmori, A., Okumura, M., Marukawa, Y., & Iwayama, M. (2003). Patent claim processing for readability – Structure analysis and term explanation. In *ACL-2003 workshop on patent corpus processing*. Sapporo: Association for Computational Linguistics.

Smith, A., & Ng, A. (2003). Web page clustering using a self-organizing map of user navigation patterns. *Decision Support Systems, 35*, 245–256.

Sperduti, A. (2001). Neural networks for adaptive processing of structured data. *Lecture Notes in Computer Science, 2130*, 5–12.

Sperduti, A., & Starita, A. (1997). Supervised neural networks for classification of structures. *IEEE Transaction on Neural Networks, 8*(3), 714–735.

Strickert, M., & Hammer, B. (2003). Unsupervised recursive sequence processing. In *Proceedings of european symposium on artificial neural networks 2003 (ESANN 2003)* (pp. 27–32).

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks, 11*(3), 586–600.

Voegtlin, T. (2002). Recursive self-organizing maps. *Neural Networks, 15*, 979–991.

Xu, P., Chang, C. H., & Paplinski, A. (2005). Self-organizing topological tree for online quantization and data clustering. *IEEE Transactions on Systems, Man, and Cybernetics, 35*(3), 515–526.