TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 09/12 pp68-76
Volume 18, Number 1, February 2013

Knowledge Map Mining of Financial Data

Wenhui Shou, Wenhui Fan*, Boyuan Liu, and Yuyang Lai

Abstract: The Knowledge Map (KM) concept, which was derived from the Fuzzy Cognitive Map (FCM), is used to describe and manage knowledge. KM provides insight into the interdependencies and uncertainties contained in the system. This paper uses a model-free method to mine KMs in historical data to analyze component stock corporations of the Shanghai Stock 50 index. The analyses use static and time-domain analyses. The results indicate that a knowledge map is useful for representing knowledge and for monitoring the health of companies. Furthermore, sudden changes of the key features of the KMs should be taken seriously by policymakers as an alarm of a crisis.

Key words: Knowledge Map (KM); data mining; health monitoring; crisis warning

1 Introduction

The Fuzzy Cognitive Map (FCM) concept is an effective approach for knowledge representation and management. FCM was first developed by Kosko in 1986, who introduced fuzzy logic into cognitive maps[1]. FCM is a graph comprising a collection of nodes, which stand for concepts or variables, and weighted arcs connecting the nodes, which represent the cause-effect relationships between them[2]. FCM has been widely applied to modeling[3], classification[4], and prediction[5].

Originally, an FCM was created according to expert perceptions, which gives poor results because it neglects the primary raw data resource so the constructed map may be subjective and incomplete. A large number of learning algorithms for FCMs have been proposed, focusing on different directions[6]. In 1994, Dickerson and Kosko[7] proposed a simple Differential Hebbian Learning (DHL) algorithm, which iteratively updated the values of the weights until they converged to certain predefined state. Schneider et al.[8] described a method for automatically creating FCMs, in which the strength of the relationship was evaluated according to the similarity between two concepts expressed as numerical vectors. Parsopoulos et al.[9] described a learning method using Particle Swarm Optimization (PSO). Stach et al.[10] utilized a Genetic Algorithm (GA) to create the FCMs, which was accurate on experimental data.

However, as the number of nodes and arcs increases, FCM construction using optimization algorithms spends much time iterating until the weight matrices reach the optimal conditions. The generated FCMs then become very complex and unreadable. To solve this problem, Allen and Marczyk[11] introduced the Knowledge Map (KM) concept, which was derived from FCM. In 2009, Marczyk[12] described a method to produce knowledge maps based on historical data and developed the commercial software OntoSpace^{TM}. The core idea of the method was to automatically extract specific knowledge from data resources to improve the map's performance and enable users to learn quickly how the information flows among the variables.

• Wenhui Shou, Wenhui Fan, and Boyuan Liu are with the Department of Automation, Tsinghua University, Beijing 100084, China. E-mail: sylvia0816@gmail.com; fanwenhui@tsinghua.edu.cn.
• Yuyang Lai is with the SOYOTEC Technologies Co., Ltd., Beijing 100081, China.
∗ To whom correspondence should be addressed.
Manuscript received: 2012-04-20; accepted: 2012-11-23

2 Knowledge Map

Knowledge map, as an extension of FCM, is a technique to describe and model knowledge on a complex system. This is also called a process map[12].

KM can be expressed as a triple $U = (V, L, C)$. $V = \{v_1, v_2, \cdots, v_n\}$ is a set of nodes that represent variables or concepts. L represents the links joining pairs of nodes, such as v_i and v_j ($v_i, v_j \in V$). The links include all the fuzzy rules used to depict the causal relationships between variables. C stands for a group of connectors which map (v_i, v_j) to (λ_{ij}, E_{ij}). λ_{ij} is the generalized cross correlation coefficient between v_i and v_j and E_{ij} is the image entropy of the corresponding scatter plot. The use of connectors can be regarded as the critical difference between KM and FCM. The correlation coefficient and entropy quantitatively depict the degree of rule fuzziness and disorganization in the system.

A typical KM is illustrated in Fig. 1. All nodes are aligned along the diagonal and the links are displayed as two vertical or horizontal segments. Different sets of nodes, such as inputs and outputs, are distinguished by different colors. The hubs that have close relationships with other nodes appear as circular shapes while the inactive nodes are shown as white squares. As with Design Structure Matrix (DSM) theory[13], segment layout avoids redundancy, which makes the map more readable.

A knowledge map provides users with a tremendous amount of information which can be summarized as[11] (1) identification of cause-effect relationships between variables, (2) visualization of how information flows within a given system, and (3) ranking of variable importance indicated by hubs and inactive nodes.

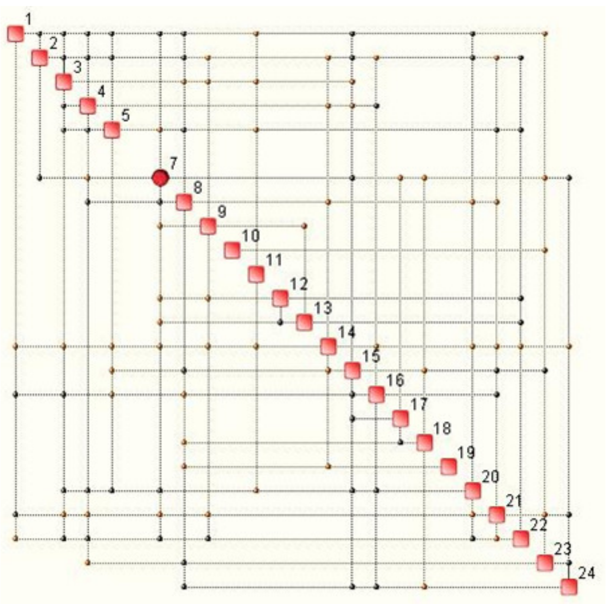

Fig. 1 Knowledge map example.

The chief task to realize KM's advantages is to obtain an objective, accurate knowledge map. As mentioned in the Introduction, maps mined from historical data are more valid and lose less information than those relying on the perceptions of experts. The mining process should be model-free without using statistical or regression methods to preserve all the information hidden in the data[12]. Unlike previous learning algorithms for FCMs, the KM mining method proposed by Marczyk eliminates the long iterative procedure and constructs a KM directly by analyzing the data resource. The following section describes how the mining process takes place.

3 Mining Process

Given a large set of numerical data

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix},$$

in which x_{kj} represents the k-th sample of the j-th variable v_j. The goal is to generate a knowledge map by mining the data matrix $\boldsymbol{X}$. KM generation process can be divided into four steps as follows.

Step 1 Scatter plot construction

All pairs of variables v_i and v_j ($1 \leqslant i \leqslant j \leqslant n$) are used to draw the scatter plots for a total of $n(n-1)/2$. One of the two key properties of each scatter plot is the generalized correlation coefficient, which ranges from 0 to 1 and is computed as[14]

$$\lambda = \sqrt{1 - \mathrm{e}^{-2I(v_i, v_j)}} \tag{1}$$

where $I(v_i, v_j)$ represents the mutual information, which is used to measure the overall linear and nonlinear dependence between v_i and v_j defined as

$$I(v_i, v_j) = \sum\sum p(v_i, v_j) \log_2 \frac{p(v_i, v_j)}{p(v_i) p(v_j)} \tag{2}$$

in which each variable is discretized and mapped into $\sqrt{N}$ intervals. In this way, the correlation is quantified to show how significantly these two variables influence each other. If λ_{ij} reaches zero, then v_i contains no information on v_j. Contrarily, $\lambda_{ij} = 1$ implies that the correlationship between v_i and v_j is so strong that one can be determined perfectly by the other.

Another property to be calculated is the entropy, which was first proposed by Shanon in 1948[15]. The image entropy of a scatter plot is

$$E_{ij} = \sum \sum p(v_i, v_j) \log_2 \frac{1}{p(v_i, v_j)} \tag{3}$$

which is a measure of the information in a plot. Disorganized plots containing large amounts of uncertainty always have large entropies. Therefore, if the image entropy of a scatter plot is relatively high while the generalized correlation coefficient is very small, then there are no significant relationships between the variables. In this situation, the plot has no significant fuzzy rules, which means no links will be created between the two variables in the final knowledge map.

Step 2 Fuzzy rule generation

The approach used here for generating the fuzzy rules from the data is essentially the same as the method used in the Association Rule Clustering System (ARCS) proposed by Lent et al.[16] However, the forms of the extracted rules are quite different. The goal here is to find how one variable reacts when another variable changes. Thus, four forms of rules are defined: IF +Delta X, THEN Delta Y; IF −Delta X, THEN Delta Y; IF +Delta Y, THEN Delta X; IF −Delta Y, THEN Delta X expressed as a quadruple. For instance, (+1,−1,+1,−1) means that if variable X increases or decreases by one unit, variable Y simultaneously changes by one unit and vice versa.

First, assume a fuzzy level (usually 3, 5 or 7) and divide the whole variable space into regions. Take 3 for example. According to the sample values, each variable is categorized into low, medium, and high group to convert the numerical vectors in the data matrix into fuzzy vectors with each sample of the original data belonging to a specific fuzzy cell.

Second, analyze all the effective scatter plots. Since the scatter plots are two-dimensional, the number of cells projected on to the plane is the square of the predefined fuzzy level. Each of these fuzzy cells should contain enough samples for the plot. Then set each cell as the coordinate origin and move rightwards (+Delta X), leftwards (−Delta X), upwards (+Delta Y), and downwards (−Delta Y) in the plot to check which cell will most likely be reached, which implies to what extent the other variable may change. If there is no dominant trend or the possibilities to arrive to several cells are equal, then no rules will be generated. Record the trend and translate the results into fuzzy rules, which are eventually noted as quadruples. An example is illustrated in Fig. 2, in which the rules are represented as (0,−1,+2,0).

Step 3 Map construction

As mentioned in Section 2, a KM is comprised of nodes, links, and connectors. First, all n variables are displayed as n nodes lying along the diagonal. The input and output variables are distinguished by the colors of the nodes.

Then establish links on the basis of the fuzzy rules generated in Step 2. If at least one quadruple exists in a scatter plot, then information is exchanged and a link is created between the variables. Otherwise, the nodes are unconnected. Each link is displayed as two vertical or horizontal segments which cross in a connector to make the map more succinct and explicit. This arrangement is referred to as the theory of DSM, which minimizes the number of segment crossings without a connector. The connectors have two critical properties as measured in Step 1 as the generalized correlation coefficient and the image entropy. They are used to illustrate how much the two variables affect each other and the amount of uncertainty in this relationship.

Finally, compute the sum of all the image entropies of those scatter plots which have quadruples and record this as the map's total entropy:

$$E_{\text{total}} = \sum_{\text{havequardruples}} E_{ij} \tag{4}$$

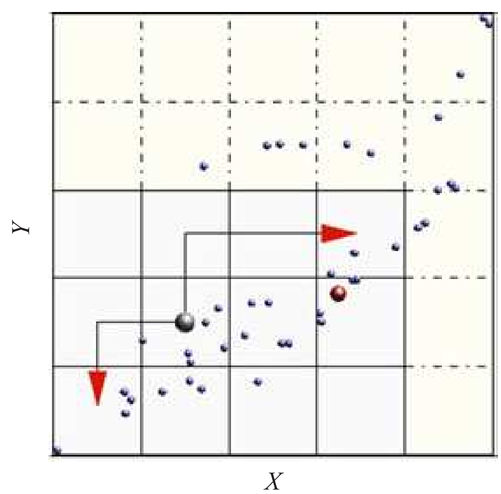

Fig. 2 Example of scatter plot with fuzzy rules.

Step 4 Hubs and inactive nodes identification

The knowledge map constructed in Step 3 is used to deduce which variable has the greatest impact on the system and which are relatively independent. A node's significance is defined by calculating how many links originate from that node. The node with the most relationships is called the hub, while nodes with no links are inactive. Both kinds of nodes are clearly represented in the map. The total number of links, l, is used to compute the map density,

$$D = \frac{2l}{q(q-1)} \tag{5}$$

in which q represents the number of active nodes. This completes the KM mining process.

4 Application

The commercial software OntoSpace$^{\text{TM}}$ was used to analyze component stock corporations in the Shanghai Stock 50 index, which includes the 50 most representative and influential stocks that are all highly fluid in the Shanghai Stock Market. OntoSpace$^{\text{TM}}$ uses this mining method and is widely used in many fields, such as product design, management, medicine, and air traffic control.

In an effort to describe the complexities of these corporations, four types of data sheets are used with balance sheets, income statements, cash flow statements, and stock market data. The companies in the financial industries have a total of 294 parameters, including 96 from balance sheets, 51 from income statements, 97 from cash flow statements, and 50 from stock market data. Companies in industries outside of finance have 238 parameters, with 75, 33, 80, and 50 in these four sheets. These historical samples cover nine years from 2002 to 2010 and have been made public on the websites. Therefore, with complete data, the data matrix will have 36 lines and 294 or 238 columns. The KMs are mined from this data for static and time-domain analyses.

4.1 Static analysis

The static analysis uses all the historical data from 2002 to 2010 to construct a KM to give a complete picture of the companies with the fuzzy levels all defined as 5. For the sake of brevity, the results of only three financial companies are presented, China Merchants Bank (CMB), China Minsheng Bank Corporation (CMBC), and Shanghai Pudong Development Bank (SPDB). These are chosen because their financial statements are more complete for the entire nine years than the other financial institutes and the data integrity has a great impact on the accuracy of the results.

Figures 3 to 5 illustrate the generated KMs for CMB, CMBC, and SPDB. Parameters from the balance sheets and cash flow statements are represented as red squares while those from the other two sheets are displayed in blue. The four key features of the maps, the number of active nodes, number of links, density, and total entropy are listed in Table 1. The number of active nodes does not differ much in the three maps. About

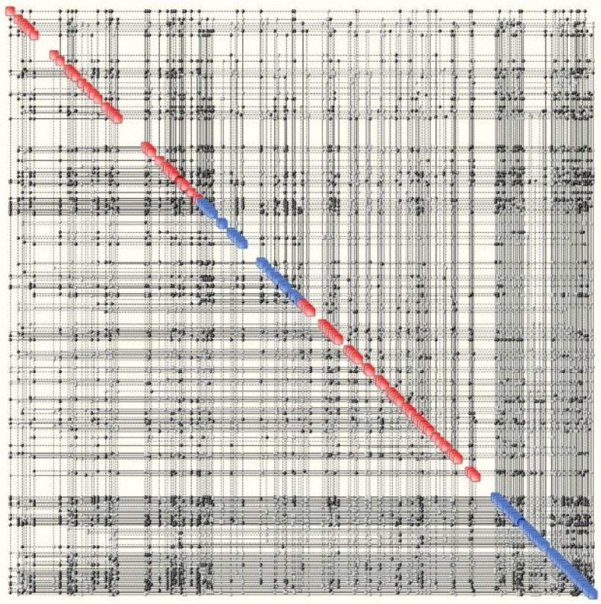

Fig. 3 CMB knowledge map.

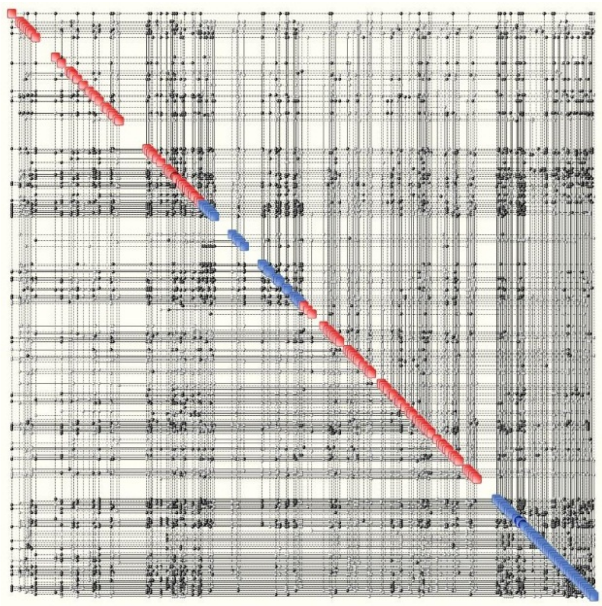

Fig. 4 CMBC knowledge map.

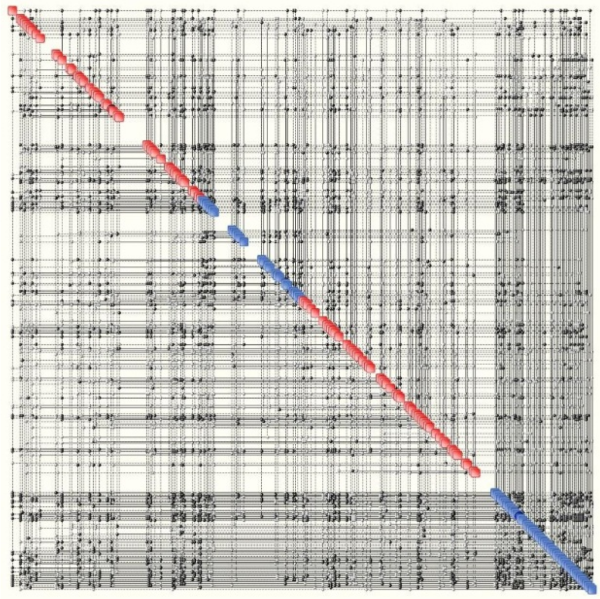

Fig. 5 SPDB knowledge map.

Table 1 Key features of the CMB, CMBC, and SPDB KMs.

Company	Number of active nodes	Number of links	Density	Total entropy
CMB	178	5702	0.36	12 181.12
CMBC	181	5479	0.34	11 543.17
SPDB	177	5490	0.35	10 624.09

39% of the 294 variables have no obvious relationships with the others, so they are sort of uncontrollable by the operators. This phenomenon is partially caused by the fact that some variables have few samples, which makes them seem to be inactive. The number of links and the density indicate that the relationships between the variables are closest in CMB, which means that it is comparatively difficult to control the company because minor disturbances in one node will be rapidly propagated to other nodes. The KM of CMB also has the highest total entropy which implies that its internal structure is more likely to collapse with the rules and patterns disappearing in a turbulent environment.

Overall, CMB has the most complicated structure with the largest amount of uncertainty in a complex system. Therefore, CMB is capable of performing more functions but its corporate health can be easily disrupted. If the environment changes and uncertainty increases, the company might become uncontrollable. The other two banks have similar complexities that are less complex than CMB with CMBC having less intimate relationships and SPDB having less uncertainty.

The data on the four sheets infer that balance sheet and stock market data contribute the most to the maps. The parameters from these two sheets correlate strongly with the other parameters to produce a large number of links. Of all the links in the CMB KM, "Total assets" and "Total liabilities" from the balance sheet have the strongest relationships with a generalized correlation coefficient of 0.91, which is quite close to 1. The scatter plot formed by these two variables shown in Fig. 6 has an image entropy of 1.63. Thus, the two variables are coordinated, which agrees with the fact that the extracted fuzzy rules are regulated as (+1,−1,+1,−1). This information then gives the interdependencies in the data, which are easily missed normally. The extracted fuzzy rules are stored as knowledge to provide guidance for understanding the companies on a deeper level.

The top 10 most important parameters in the three companies are listed in Tables 2 to 4 as measured by the number of links originating from the nodes. For example, S4-v14, the 14th variable in the 4th data set, is the most important parameter in CMB. Although the hubs are not identical, the three companies are all related to a large number of important nodes, including "Total assets", "Total liabilities", and "Deposits". Consequently, these parameters are quite influential and worthy of close attention by the policymakers. Eight out of ten important variables are from the balance sheet, which leads to the conclusion that the balance sheet plays a dominant role in the system. The hubs in the KMs explicitly identify the most vulnerable positions and the company should be careful about changes to these which might provoke collapse of the entire system.

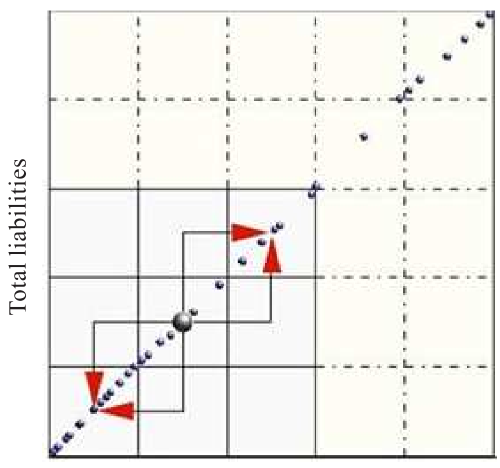

Fig. 6 Example of CMB scatter plot.

Table 2 Top 10 most important parameters in CMB.

Variable	Variable name	Number of links
S4-v14	Listed tradable shares	124
S1-v86	Surplus reserves	123
S1-v45	Total assets	122
S1-v51	Deposits	122
S1-v96	Total liabilities and shareholder equity	122
S1-v82	Total liabilities	121
S1-v26	Loans and advances	116
S1-v34	Fixed assets	116
S1-v71	Interest payable	116
S4-v43	ROE weighted	115

Table 3 Top 10 most important parameters in CMBC.

Variable	Variable name	Number of links
S1-v83	Share capital	119
S1-v82	Total liabilities	118
S1-v70	Taxes expense payable	117
S1-v45	Total assets	116
S1-v88	Retained earnings	116
S1-v96	Total liabilities and shareholder equity	116
S3-v61	Net profits	116
S1-v51	Deposits	115
S1-v71	Interest payable	115
S1-v1	Cash deposits and required reserve in central bank	114

Table 4 Top 10 most important parameters in SPDB.

Variable	Variable name	Number of links
S1-v51	Deposits	129
S1-v82	Total liabilities	120
S1-v45	Total assets	119
S1-v96	Total liabilities and shareholder equity	119
S1-v87	Contingency reserves	117
S4-v13	Tradable shares	117
S3-v76	Decrease (less: Increase) in operating receivables	115
S1-v86	Surplus reserves	114
S4-v15	Multiplying cumulative adjusting factor for price	113
S1-v1	Cash deposits and required reserve in central bank	112

4.2 Time-domain analysis

In the time-domain analysis, the samples were split into several continuous periods or windows. Only one window of samples was examined in each step and the windows were overlapped so that the results of each step did not differ significantly. The window width was defined as 12 and the overlap step was 11. Financial statements and stock market data for 12 quarters was used to mine a knowledge map. A total of 25 maps were generated and used to track the company's performance and to monitor the system health in terms of structure and uncertainty.

This study uses three typical corporations, SAIC, Sinopec, and Gemdale, in the three non-financial industries of manufacturing, oil extraction, and real estate. Their samples were complete from 2002 to 2010, which makes the analysis results more credible.

Figure 7 illustrates the evolution of the four key features of the knowledge maps constructed in seven steps corresponding to seven years. The first three describe the structural changes of the companies from different aspects while the last one presents the shifting trends of the uncertainty.

As reflected in Fig. 7a, SAIC and Sinopec experienced rapid growth in the number of active nodes in 2006-2007, with the increase in Gemdale not as dramatic. Figure 7b shows that in 2007, the number of links all surged higher in the three companies. The Sinopec and Gemdale map densities reached their highest points in the same year according to Fig. 7c, while the density of the SAIC map remained relatively stable. These figures show that the structures of the three corporations became uniformly more complex in 2007 as seen in the growth in the numbers of links. The total entropy trend chart in Fig. 7d further indicates that the uncertainties of all three corporations, especially SAIC, increased up to 2007.

The complex structure coupled with uncertainty can cause difficulties in controlling and managing the companies. The sudden change in the number of links and the total entropy should be taken seriously as a sign of crisis. In an attempt to confirm whether the increase in 2007 was a special local case, comprehensive financial and stock data was analyzed for all the component stock corporations of the Shanghai Stock 50 index for the period 2002-2010. A total of 12 corporations satisfied the conditions, including the three banks analyzed in static analysis. The number of links and the total entropy in their maps in 2006 and 2007 are listed in Table 5 with the growth rate for each corporation. 75% of the companies experienced a sharp rise in the number of links during 2007 with an average growth rate of 13%. In addition, the total entropy

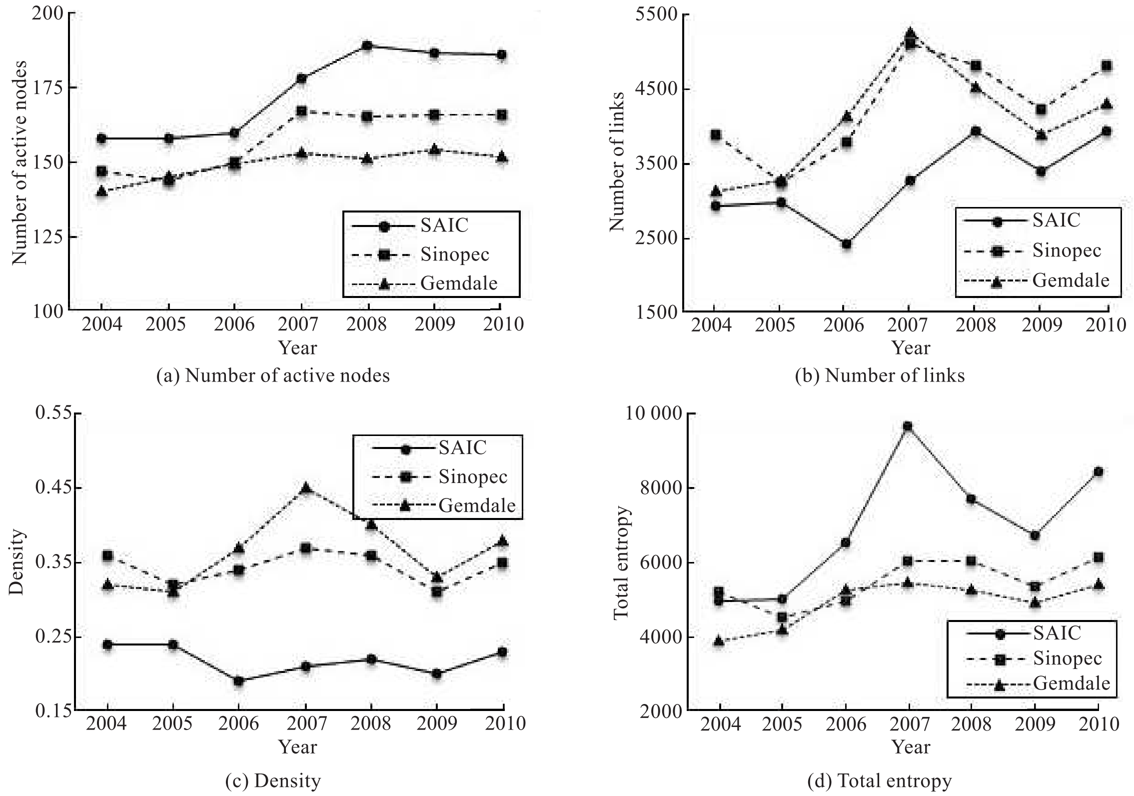

Fig. 7 Evolution of four key features of the KMs of SAIC, Sinopec, and Gemdale.

Table 5 Number of links and total entropy of the KMs for 12 corporations in 2006 and 2007.

Company	Number of links			Total entropy		
	2006	2007	Growth rate (%)	2006	2007	Growth rate (%)
CMB	2223	2045	−8	4885	4651	−5
CMBC	2444	2661	9	6019	6242	4
SPDB	2417	2406	0	5495	5672	3
SAIC	2425	3266	35	6522	9659	48
Sinopec	3796	5128	35	4973	6045	22
Gemdale	4131	5253	27	5239	5428	4
BaoSteel	6791	5583	−18	7743	6466	−16
TBEA	4074	5233	28	5392	5886	9
Baotou	5116	6897	35	5702	7386	30
Yanzhou	3315	3662	10	4043	4331	7
Jiangxi Copper	4807	5007	4	5646	5482	−3
Conch Cement	3603	4095	14	4369	4700	8
Average	3762	4270	13	5502	5996	9

increased by 9% on average, with increases in 9 of the 12 companies.

These results suggest that the critical properties of the KMs generally increased in 2007. Many reasons may have led to this situation, with the global economic environment playing a key role. As the global economy became unstable, uncertainty will inevitably increase in the companies and the functionality tends to be more complicated, which is directly associated with changes in the company structures. The main characteristics of the maps reflect not only the operating conditions of these systems but also the state of the entire environment to some extent. The global financial crisis began in 2008. Therefore, the sudden growth in the number of links and the total entropy, which correspond to the company structure and uncertainty, was an early warning signal of the 2008 financial crisis. If the anomalous changes during 2006-2007 were promptly reported to the managers, they could have prepared better for the financial crisis and taken appropriate measures ahead of the situation to minimize damage and avoid losses. Thus, knowledge map mining is an effective alarm tool.

Above all, the time-domain analysis allows managers to monitor the operating condition of the corporation in terms of structure and uncertainty. A steep rise in the complexity of the knowledge map indicates that the company is encountering the turbulence and is likely to collapse since the interconnections within the system tend to be complex and fragile.

5 Conclusions

This paper describes a mining method to construct knowledge maps utilizing historical data without the intervention of domain experts. The software OntoSpace^{TM} is used to apply the method to analyze component stock corporations of the Shanghai Stock 50 index. The static analysis results show that the KM is capable of discovering the structure of the examined systems given through fuzzy rules, interdependencies, hubs, and inactive nodes, and can identify the uncertainty as measured by the entropy. Time-domain analysis reveals the evolution of the main features of the knowledge maps, which can be used by policymakers to monitor the company health. The results demonstrate that the mining of knowledge maps from company data is remarkably valuable and that tracking of the main properties of the KMs can effectively indicate crises, which is not possible by conventional risk rating methods.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 60874066) and the National High-Tech Research and Development (863) Program of China (No. 2009AA110302).

References

[1] B. Kosko, Fuzzy cognitive maps, *International Journal of Man-Machine Studies*, vol. 24, no. 3, pp. 65-75, 1986.

[2] J. Aguilar, A survey about fuzzy cognitive maps papers, *International Journal of Computational Cognition*, vol. 3, no. 2, pp. 27-33, 2005.

[3] L. Rodriguez-Repiso, R. Setchi, and J. L. Salmeron, Modelling IT projects success with fuzzy cognitive maps, *Expert Systems with Applications*, vol. 32, no. 2, pp. 543-559, 2007.

[4] Z. Peng, B. Yang, C. Liu, Z. Tang, and J. Yang, Research on one fuzzy cognitive map classifier, (in Chinese), *Application Research of Computers*, vol. 26, no. 5, pp. 1757-1759, 2009.

[5] T. Hong and I. Han, Knowledge-based data mining of news information on the Internet using cognitive maps and neural networks, *Expert Systems with Applications*, vol. 23, no. 1, pp. 1-8, 2002.

[6] E. I. Papageorgiou, Learning algorithms for fuzzy cognitive mapsła review study, *IEEE Trans on Systems, Man and Cybernetics*, vol. 42, no. 2, pp. 150-163, 2012.

[7] J. A. Dickerson and B. Kosko, Virtual worlds as fuzzy cognitive maps, *Presence*, vol. 3, no. 2, pp. 173-189, 1994.

[8] M. Schneider, E. Shnaider, A. Kandel, and G. Chew, Constructing fuzzy cognitive maps, in *Proc. 1995 IEEE International Conference on Fuzzy Systems*, Yokohama, Japan, 1995, pp. 2281-2288.

[9] K. E. Parsopoulos, E. I. Papageorgiou, P. P. Groumpos, and M. N. Vrahatis, A first study of fuzzy cognitive maps learning using particle swarm optimization, in *Proc. 2003 Congress on Evolutionary Computation*, 2003, pp. 1440-1447.

[10] W. Stach, L. Kurgan, W. Pedrycz, and M. Reformat, Learning fuzzy cognitive maps with required precision using genetic algorithm approach, *Electronics Letters*, vol. 40, no. 24, pp. 1519-1520, 2004.

[11] G. Allen and J. Marczyk, Tutorial on complexity management for decision-making, http://www.e-education.psu.edu/drupal6/files/sgam/Tutorial%20on%20 Complexity%20Management%20for%20Decision-Making.pdf, 2012.

[12] J. Marczyk, *A New Theory of Risk And Rating*, Trento: Editrice Uni Service, 2009.

[13] D. V. Steward, The design structure matrix: A method for managing the design of complex systems, *IEEE Transactions on Engineering Management*, vol. EM-28, no. 3, pp.71-74, 1981.

[14] S. Aumonier, Generalized correlation power analysis, in *Proc. ECRYPT Workshop on Tools For Cryptanalysis*, Krakw, Poland, 2007.

[15] C. E. Shannon, A mathematical theory of communication, *Bell System Technical Journal*, vol. 27, pp.379-423, 1948.

[16] B. Lent, A. Swami, and J. Widom, Clustering association rules, in *Proc. 13th International Conference on Data Engineering*, Birmingham, England, 1997, pp. 220-231.

Wenhui Shou, born in 1990, is a master degree candidate in Tsinghua University now. She received her BEng degree in Beijing University of Posts and Telecommunications in 2010. She has published three papers in two international conferences indexed by EI Compendex and a Chinese core journal since 2011. Her research interests include data mining and artificial intelligence.

Wenhui Fan, born in 1966, is currently an associate professor in the Department of Automation, National CIMS Engineering Research Center, Tsinghua University, Beijing, China. He received his PhD degree in Zhejiang University in 1998. Research interests include multidisciplinary collaborative design, simulation and optimization.

Boyuan Liu, born in 1985, is a PhD candidate in Tsinghua University now. He received his BEng degree and MEng degree in Harbin Institute of Technology. His research interests include data mining and system engineering.

Yuyang Lai, the founder and CEO of SOYOTEC LIMITED, has over ten years of experience of complexity management, computer aided engineering and design optimization. He received his BEng degree and MEng degree in the Department of Thermal Engineering, Tsinghua University in 2000 and 2002. He currently works with global companies including aerospace, automotive and off-shore industries, helping them measure and reduce their business complexity and optimize the performance and reliability of their products.