行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※適用於科學資料的資料管理與分析系統之研究(II)※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：■個別型計畫　　□整合型計畫
計畫編號：NSC　90-2213-E-009-130
執行期間：90 年　8　月　1　日至 91 年　7　月 31　　日

計畫主持人：梁　婷
共同主持人：

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

執行單位：國立交通大學資訊科學系

中　華　民　國　　91　年　10　月　30　日

行政院國家科學委員會專題研究計畫成果報告

適用於科學資料的資料管理與分析系統之研究(II)

Study of A Data Management and Analysis System for Scientific Data (II)

計畫編號：NSC 90-2213-E-009-130

執行期限：90 年 8 月 1 日至 91 年 7 月 31 日

主持人：梁　婷　　國立交通大學資訊科學系

計畫參與人員：洪炎東、時繼弘、王建邦 國立交通大學資訊科學系

一. 中文摘要

隨著資料庫技術的發展，各式各樣的數據大量的產生。如何進一步將數據轉換成有用的信息是有非常迫切的需要。因此在本計畫中我們將針對大量計算流體力學數據提出並製作一個資料歸納系統。此系統包含三個主要模組：區域特徵萃取、分類機制、和廣域特徵探勘機制。其中區域特徵萃取已在先前計畫中完成視覺特徵的萃取。在本計畫中我們將進行從流場資料中萃取出可用的統計特徵，並進一步藉由分類機制的應用將流場資料作廣域的彙整，以提供使用者一個全觀的了解。同時利用探勘技術萃取出重要的廣域特徵，進而發掘出資料間特徵和群體間的依存關係。我們希望藉由本計畫的執行，不僅在知識探勘技術上有進一步探討與應用，也能經由系統的實做提供流力研究者一個好用的分析與管理工具。

關鍵詞： 資料管理、歸納、特徵萃取、分類、資料探勘、計算流體力學.

Abstract

With the advent of database technology, various communities have accumulated increasingly large amount of data. Data from various activities need further analysis to transform it into useful information. In this project we plan to implement a summarization system designed for flow field data. The system will contain three main components: local feature extractor, cluster generator and global feature finder. The local feature extractor has successful extracted the visual features with image processing techniques in our previous project so far. In this project those useful statistical features will be further extracted from flow field data by the extractor. Meanwhile the cluster generator and the global feature finder will be implemented by data mining techniques to provide users with an overall understanding of data content, such as global features and the dependencies between features and clusters. The implementation of this project will benefit both the information scientist in the context of knowledge discovery and at the same time provide an efficient flow data management and analysis tool for researchers in the computational fluid dynamics community.

Keywords: data management, summarization, feature extraction, clustering, data mining, computational fluid dynamics.

二. 緣由與目的

In our previous research project (supported by National Science Council, No. NSC

89-2213-E009-176), the issues of features extraction, key frame extraction, frame indexing, and user interface of a data management and analysis system for scientific data had been addressed. The major features such as vortex center, stagnation point, separation point, and vortex size have been successfully identified. These features are local to each time step within the whole data set. But there are important global features, such as similarity between flow fields at different time steps, the relationship among features, ⋯ , etc., that should be addressed. To discover these global features and provide a user with an overall view of data, a summarization system is certainly demanded and is an indispensable part of a data management and analysis system.

In this project, there are two major issues that will be dealt with in designing the summarization system. The first issue is the design of clustering algorithms suitable for grouping flow field data into clusters with respect to their content. The second is the design of data miner so that these frequent patterns within flow field sequences and association rules between feature attributes can be effectively.

In the past years, there are various kinds of global clustering algorithms proposed and investigated in literature [2, 3, 4, 5, 6]. One popular algorithm is the K-means algorithm [6] in which clustering is based on minimization of the overall sum of squared errors between each pattern and corresponding cluster centers. This approach is easy to be implemented and is practical to deal with clustering of large volume of data sets. However, it suffers from several drawbacks. First the user has to know a priori the number of clusters presented in the data. Second, the objective function is not convex, and hence, it may stick in local optimal solution. Finally, the performance of this algorithm depends on the choice of the initial cluster centers. To avoid sticking in local optimal solution, many optimization methods are proposed, namely simulated annealing (SA) [5], genetic algorithm (GA) [4], and evolutionary programming (EP) [3]. SA is a kind of sequential search algorithm for finding optimal solutions. The quality of the clustering results can be controlled by the number of iterations. GA and EP are both parallel search algorithms. They start with P different paths and always try to generate new paths, which are better than the current paths. In [2] a clustering algorithm based on semi-Hausdorff distance measure is proposed, and it transforms clustering problem to a covering problem. In this project, a clustering algorithm based on semi-Hausdorff distance measure is implemented. The benefit of our implemented clustering algorithm is that it scans the whole data set only once. This is especially important for dealing with large amount of flow field data.

Except to the clusters which can be treated as an overview of data set, other issues, such as the relationship between feature and clusters, should also be addressed in designing the summarization module of management and analysis system. In order to investigate such relationship, a feature finder to discover the significant features from clusters will be investigated in this proposal. In fact, selecting a right feature set not only can improve subsequent classification accuracy but also can reduce the running time of predictive algorithms and lead to simpler, more understandable models [8].

One approach to generate significant features is statistical χ^2 method, with which a fixed number of significant features can be extracted to represent a cluster centroid for textual classification. Another approach based on a genetic algorithm was implemented in our previous work [7]. The experimental results from [7] show that the genetic approach is indeed capable to yield a better set of discriminating features in terms of higher classification accuracy rate than the usual statistical method. On the other hand, a scalable feature mining for sequential data was proposed by [8]. This approach is based on clustering algorithm and sequence mining and it was shown to be efficient to select features from large data set.

In this project, the feature miner is designed to capture the frequently occurring patterns related to a sequence. Apriori algorithm [1] is commonly used to mine frequent patterns in sequences and is also implemented in our data miner module to mine the frequent patterns in flow field data sequences. To understand the relationships between features of flow fields at different time instants, the data-mining tool, DBMiner, is used and several association rules are expected to be mined.

It is our purpose that, throughout the implementation of this project, those possible new and better ways to manage and analyze large flow data sets from computational fluid dynamics can be explored. On the other hand, this project will benefit both the information scientist in the context of knowledge discovery and at the same time help develop a good data analysis system for the fluid dynamist to better deal with the flow data.

三. 結果與討論

The proposed system was implemented in Windows 2000, CPU P3 (Intel), and a platform independent language IDL (interactive data language) from Research System Inc. Figure 1 shows the system architecture. The testing flow field data were obtained from our collaborator Professor Robert Hwang from Naval Structure and Ocean Engineering Department, National Taiwan University.

Identification of shots is first done with the identification of shot types which are defined by the changing degree between adjacent frames. Then a local clustering method based on maximum-block-difference is implemented to identify each shot from sequence of frames. Figure 2 shows one local clustering result example from dataset 1. Twelve frames shown in figure 2 are identified as two shots, frame 697 to 702 is one shot, and frame 703 to 708 is another one. Figure 4 shows the shot viewing window of the system.

On the other hand, a global clustering algorithm based on semi-Hausdorff distance measure is implemented in the proposed system to extract the main data types of a set of flow field data. The benefit of our implemented global clustering algorithm is that it only scans whole data set once. This is especially important for dealing with large amount of flow field data containing more than 10^6 time instants data. Figure 3 show a global clustering result example. In figure 3, several frames selected from the 56^{th} cluster of testing data set are shown. Figure 5 shows the cluster viewing window, and the first frame of the 56^{th} cluster is shown on the draw area of that window.

To discover the frequent patterns within a

flow field data sequence, we use the cluster identifier of each frame to represent the data sequence, and this sequence is generalized to a compact list. The frequent patterns within a flow field sequence are successfully extracted by applying the Apriori algorithm to the compact list. Table 1 lists four mined two-item frequent patterns and the corresponding segments where these patterns occur.

In addition the relationships among features are also investigated and extracted by association rule mining. To mine association rules, numerical features are transformed into nominal data types by the mining tool DBMiner. Table 2 lists the four mined association rules among vortex number and separation point number from dataset 1.

四. 成果自評

In this project a practical vortex information system with helpful summarization tool is well constructed. The system was verified by some hydrodynamists to be very useful for them to study and manage their data. Throughout the implementation of the system, some empirical mining techniques are investigated and novel methods are also proposed. In the course of this project students are trained to actually implement algorithms for particular usage. The applicant believes that such a system may even shed light on knowledge discovery. In the end we appreciate financial supports from National Science Council to the extension of this project.

參考文獻

[1] Agrawal, R., Imielinski, T., and Swami, A. (1993), "Mining Association Rules Between Sets of Items in Large Databases," Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 207-216.

[2] H. S. Chang, S. Sull, and S. U. Lee (1999), "Efficient Video Indexing Scheme for Content-Based Retrieval", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No. 8, pp. 1269-1279.

[3] D. B. Fogel (1994), "An Introduction to Simulated Evolutionary Optimization," IEEE Transactions on Neural Networks, Vol. 5, Issue 1, pp. 3-14.

[4] D. E. Goldberg (1989), Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading, MA, USA.

[5] S. Z. Selim, and K. A. Sultan (1991), "A Simulated Annealing Algorithm for the Clustering Problem," Pattern Recognition, Vol. 24, Issue 10, pp. 1003-1008.

[6] J. Tou, and R. Gonzalez (1974), Pattern Recognition Principles, Addison-Wesley, Reading, MA, USA.

[7] Tyne Liang and C. H. Kuo, (2000), "A genetic approach to class descriptor generation," Proceedings of the 2000 International Computer Symposium, Workshop on Artificial Intelligence, pp. 134-141.

[8] N Lesh, M. J. Zaki, M. Ogihara, (2000), "Scalable Feature Mining for Sequential Data," IEEE Intelligent Systems, pp. 48-56.

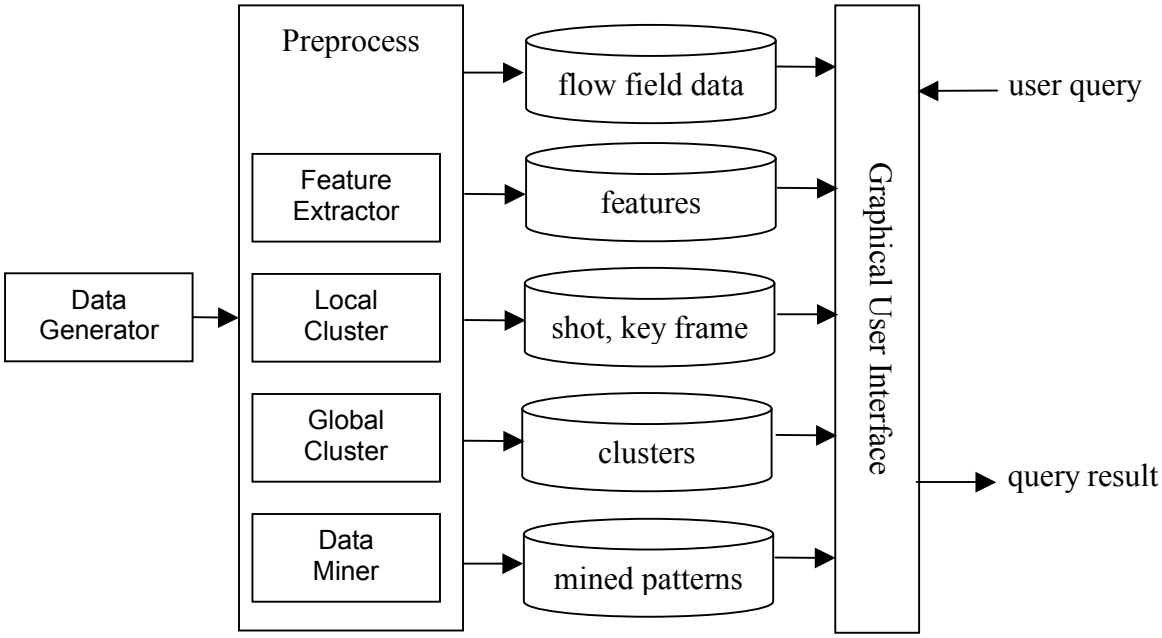

figure 1. The system architecture.

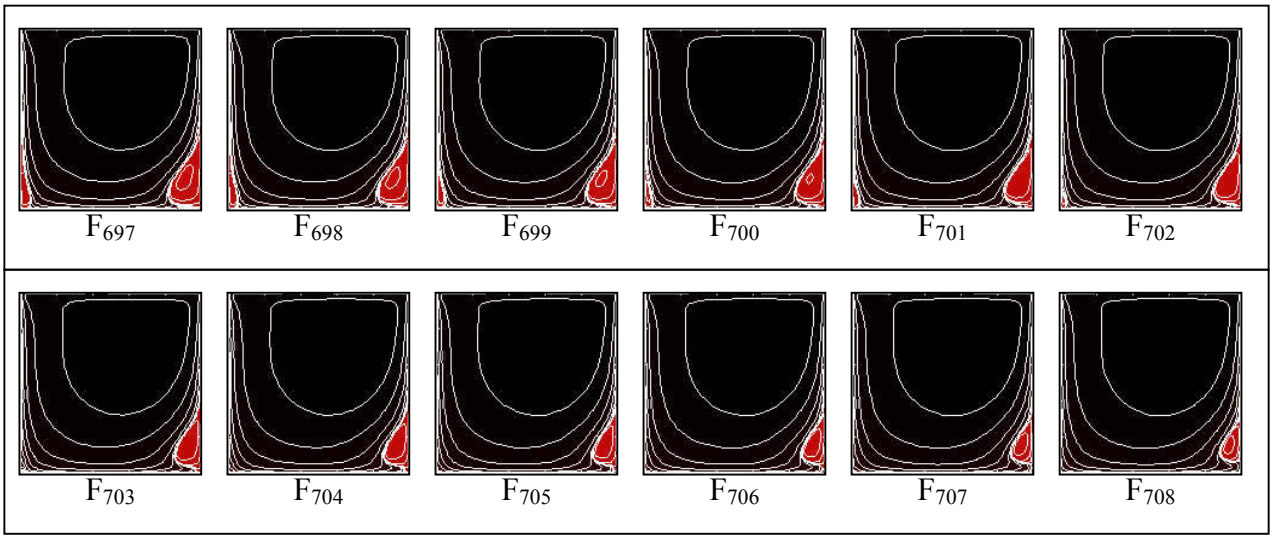

figure 2: A local clustering result example.

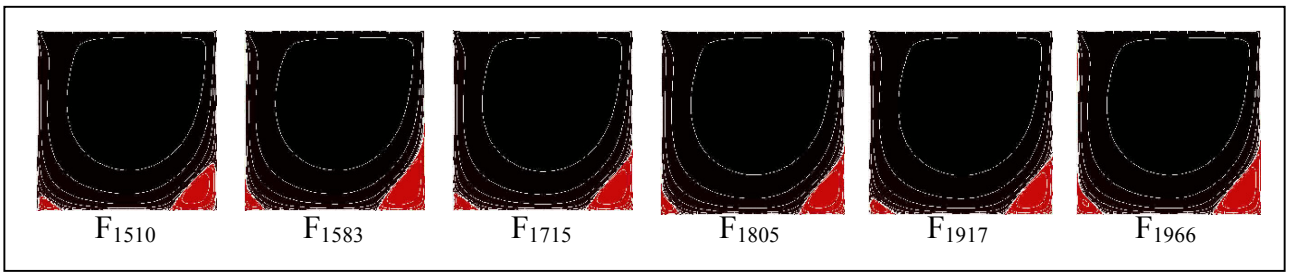

figure 3: A global clustering result example.

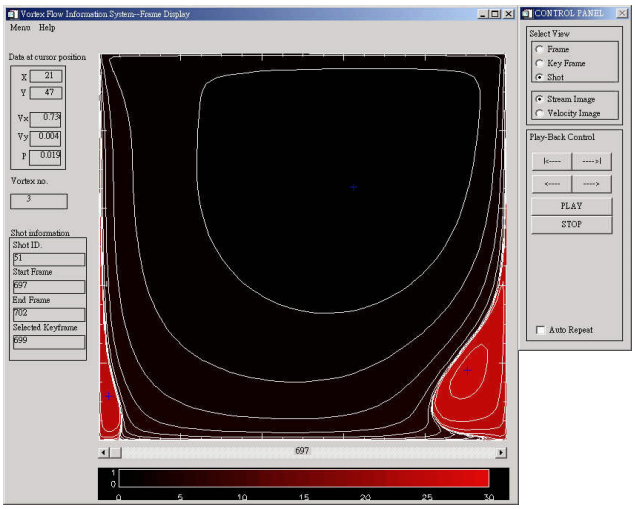

figure 4. The shot viewing window.

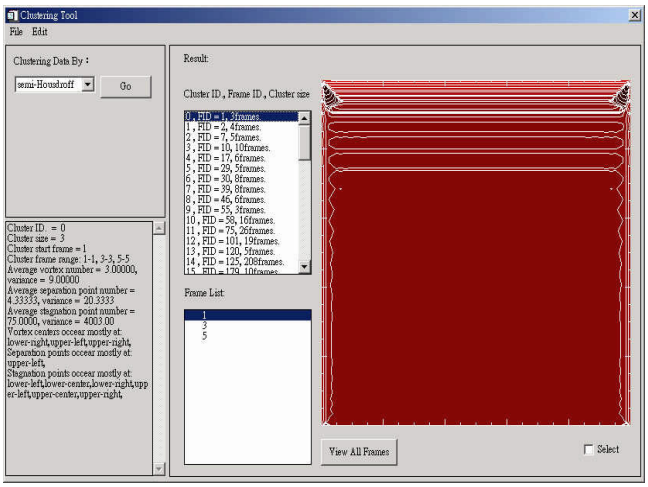

figure 5. The cluster viewing window.

Table 1. Frequent pattern list.

Pattern	Occurs at:
“0,1”	F_1~F_6
“53,54”	F_{923}~F_{1025}, F_{1114}~F_{1228}, F_{1310}~F_{1430}
“54,51”	F_{995}~F_{1097}, F_{1206}~F_{1297}, F_{1385}~F_{1457}
“55,56”	F_{1458}~F_{1598}, F_{1623}~F_{1809}, F_{1831}~F_{2000}

Table 2. Association Rule list.

Rule	Support	Confidence
2 vortex => 2 sep. pt.	19.25	99.483
3 vortex => 4 sep. pt.	20.5	92.135
4 vortex => 6 sep. pt.	16.9	81.446
5 vortex => 8 sep. pt.	14.45	87.895