TCMiner: A High Performance Data Mining System for Multi-dimensional Data Analysis of Traditional Chinese Medicine Prescriptions[1]

LI Chuan, TANG Changjie, PENG Jing, HU Jianjun

The Data Base and Knowledge Engineering Lab (DBKE)

Computer School of Sichuan University
{lichuan, tangchangjie, pengjing, hujianjun}@cs.scu.edu.cn

Jiang Yongguang, Liu Juan

Chengdu University of Traditional Chinese Medicine

{cdtcm, liujuan0}@163.com

Abstract. This paper introduces the architecture and algorithms of TCMiner: a high performance data mining system for multi-dimensional data analysis of Traditional Chinese Medicine prescriptions. The system has the following competing advantages: (1) High Performance (2) Multi-dimensional Data Analysis Capability (3) High Flexibility (4) Powerful Interoperability (5) Special Optimization for TCM. This data mining system can work as a powerful assistant for TCM experts by conducting Traditional Chinese Medicine Data Mining such as Computer Aided Medicine Pairing Analysis, Medicine Syndrome Correlation, Quality and Flavor Trend Analysis, and Principal Components Analysis and Prescriptions Reduction etc.

1 Introduction

Traditional Chinese medicine (TCM) has a long therapeutic history of thousands of years and the therapeutic value of which, especially on chronic diseases, has been winning wider and wider acknowledgement in the World [1]. In addition, the TCM seems to have made enormous strides forward after China's entry into the World Trade Organization (WTO), as large sums of capital investment become available to spur technical innovation. The World Health Organization (WHO) has also been keen to pursue the development of TCM in recent years.

[1]This Paper was supported by Grant of National Science Foundation of China (60073046), Specialized Research Fund for Doctoral Program by the Ministry of Education (SRFDP 20020610007), and the grant from the State Administration of Traditional Chinese Medicine (SATCM 2003JP40). LI Chuan, PENG Jing, HU Jianjun are Ph. D Candidates at DB&KE Lab, Sichuan University. Jiang Yongguang is a Professor at Chengdu University of Traditional Chinese Medicine. And TANG Changjie is the associate author.

S. Wang et .al. (Eds.) ER Workshops. 2004 LNCS3289,pp246-257,2004-11-22
Springer-Verlag, Berling Heidelberg 2004

However, despite its existence and continued use over many centuries, and its popularity and extensive use during the last decades, its chemical background and formula synergic effects are still a mystery at least in theoretical sense because of its complex physiochemical [2]. Newly developed techniques, such as data mining or knowledge discovery in database (KDD) which aim at discovering interesting patterns or knowledge from large scale of data by non-trivial approaches, provide us with a very promising means and a hopeful opportunity to do research on the TCM data.

Designing a data mining system for analysis of TCM data has been considered by both the TCM and data mining trade for quite a long period but due to lack of mutual understanding and the true complexity of the problem itself the work seems complicated and challenging. The Data Base and Knowledge Engineering Lab (DBKE) at Computer School of Sichuan University in Chengdu has been working on TCMiner (Traditional Chinese Medicine Miner) in collaboration with Chengdu University of Traditional Chinese Medicine under a research grant from State Administration of Traditional Chinese Medicine (SATCM) to investigate new methods of multi-dimensional data analysis of Traditional Chinese Medicine prescriptions. This system can provide knowledge discovery and data mining capabilities for TCM data values as well as for categorical items, revealing the regularities of TCM paring and indicating the relationship of prescriptions, medicine and syndromes.

TCMiner has the following distinguishing features: (1) High Performance: Algorithms implemented in the system are all the leading algorithms in their respective domains. (2) Multi-dimensional Data Analysis Capabilities: The system has multi-dimensional analysis capabilities. E.g. it can discover Medicine Pairing regularities in consideration of digitalized Quality, Flavor, Channel tropism etc. in form of multi-dimensional frequent patterns or association rules (3) High Flexibility: In order to realize the highest flexibility of the system to meet the different requirements of different users, multiple engines were adopted. (4) Powerful Interoperability: By the highly interactive visual and friendly user interface, the system has wonderful interoperability. (5) Special Optimization for TCM: TCMiner optimizes the system architecture and algorithms according to characteristics of TCM data in the following aspects: (a) Standardization of discrete attributes such as Quality, Flavor, Channel tropism (b) TCM Knowledge Base Constructions for heuristic search (c) The system implements General Trend Evaluation and Contrast of Quality, Flavor, and Channel tropism, which proves to be of high value in general analysis of TCM prescriptions. (d) Construction of TCM Data Warehouse (e) Multi-dimensional Prescriptions Structure Analysis.

The remaining of the paper is organized as follows. Section 2 introduces the methodological knowledge in TCM data analysis. Section 3 covers the special designing issues concerning the particular TCM data process and system implementation. Section 4 details the system architecture of TCMiner. Section 5 presents the high-level performance of the fundamental algorithms based on which TCMiner engines are implemented. Section 6 exhibits the system interface. Section 7 discusses the probable future directions of TCMiner's later versions. And section 8 concludes the paper.

2 Methodology of TCM Data Analysis

TCM therapy theory is a very broad and deeply philosophy in that it contains not only the theoretical explanations of how TCM works but also comprises the particular prescriptions, medicines, herbs and laws. Traditional Chinese Medicine Prescriptions are the original records of detailed procedures and ways in which theses herbs and medicines are put to use. What

implied therein are the Paring Regularities of TCM, actually represented in form of the correlations among Prescriptions, Medicines, and Syndromes. The major objective of the TCM Prescription Analysis Research is to reveal these corresponding associations by means of the latest data processing and analysis technology—Data Mining.

2.1 A glance at TCM prescription

Before we go into the complicated technical details, let's have a glance at a typical TCM prescription shown below (extracted from TCM professional materials). The prescription consists of two medicines/herbs: Corktree bark and Atractylodes rhizome with their corresponding Quality, Flavor, Channel tropism, Function and Dose shown to the right. For detailed explanation of the terms and notations therein, please refer to reference [1].

	1st layer	2nd layer	3rd layer	
1st level	Components:	Corktree bark	Quality	Cold
			Flavor	Bitter
			Channel tropism	Kidney/Bladder
			Function	Resolving heat/ Eliminating dampness/ Consolidating Yin
			Dose	15g
		Atractylodes rhizome	Quality	Warm
			Flavor	Spicy and Bitter
			Channel tropism	Spleen/Stomach
			Function	Eliminating dampness
			Dose	15g
2nd level	Function:	Resolving heat/ Eliminating dampness		
	Indications:	Pathogenesis	downward flow of damp-heat	
		Syndrome	Flaccidity syndrome/Arthragia syndrome /Dermatophytosis /Leukorrhea /Eczema	
3rd level		Pulse	soft and floating pulse	
		⋮		
	Add/Reduce:		add Chaenomeles fruit /Dioscorea septemloba	
4th level	Derived Prescription	San Miao Powder /Si Miao Pill/Qi wei cang bai Powder/Qing re sheng shi Decoction		

Fig. 1. A sample TCM prescription

2.2 Technological routines

Since Prescriptions Paring Regularities contain two indispensable factors: Medicine Paring Regularities and Prescription-Syndrome Matching Regularities, through continuous and ardent discussions with TCM experts, we are resolved to adopt the following general line: Start with prescription structure analysis and medicine syndrome analysis concentrating on the medicine components data, and then conduct research on prescription - syndrome matching regularities regarding "Prescription-Medicine-Syndrome" as a whole. The research routine is outlined as follows:

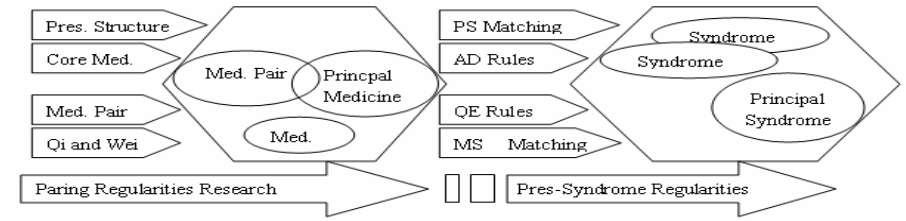

Fig. 2. The general technique routine of TCM prescription analysis

As the first step, we investigate the prescription structure analysis method with the following sub-objectives and solutions: (1) basic prescription finding: first, track down the principal medicine components in the prescriptions; second, sort and rate these medicines according to their importance; third, cored with principal medicine, summarize the structure of fundamental and derived prescriptions; finally, propose the initial formation of a basic prescription to the TCM professionals for evaluation (2) medicines group selection: first, try to find a group of medicines which are often used together in the same kind of prescriptions; second, refine the medicine pairs and groups through testing samples; third, ascertain the medicine pairs and groups through special "medicine-syndrome" matching (3) Medicine Syndrome correlation analysis: the law of the medicine usage for a certain disease or corresponding syndromes in form of multi-dimensional representation; the different medicine usage laws for different parts of the body with different Quality and Flavor. The analysis figure is sketched below, where arrows represent data flow (due to space limitation, detailed interpretations are not given here):

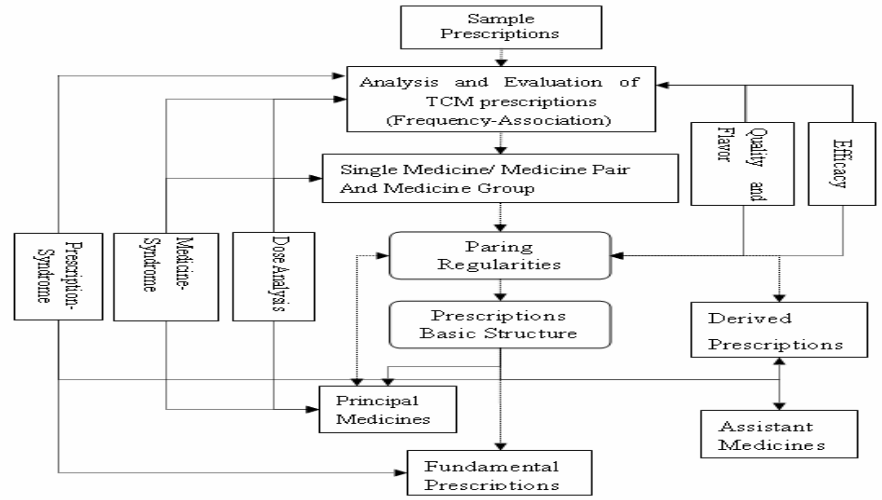

Fig. 3. Technique routine of prescription analysis

In the second stage, we concentrate on the prescription-syndrome correlation analysis with the following summarized analysis contents: (1) Organization characteristics analysis of categorical prescriptions or related syndromes including: (a) the Recognition of major syndromes and by syndromes (single syndrome, syndromes group, and syndromes herd) (b) the analysis and judge of combining patterns cored by principal syndrome. (2) The Analysis of Syndrome Data: (a) the analysis of Cold/Hot Trend, Weak/Strong Trend, Corresponding Body Parts, and Pathogenesiss and characteristics of the Syndromes; (b) The correlations of related syndromes; (c) the significance and frequency of the syndromes. (3) The Analysis of Syndromes through prescriptions and medicines: (a) the deductive and verifying analysis through the Quality, Flavor, Channel tropism, function and efficacy prescriptions and medicines; (b)the verification and deduction through the associations of the efficacy, clinical effects and Pathogenesiss of prescriptions, esp. basic prescriptions. (4) The methods of construction of the syndrome knowledge base. The syndrome data analysis threads are summarized into a figure as follows, where arrows represent data flow (due to space limitation, detailed interpretations are not given):

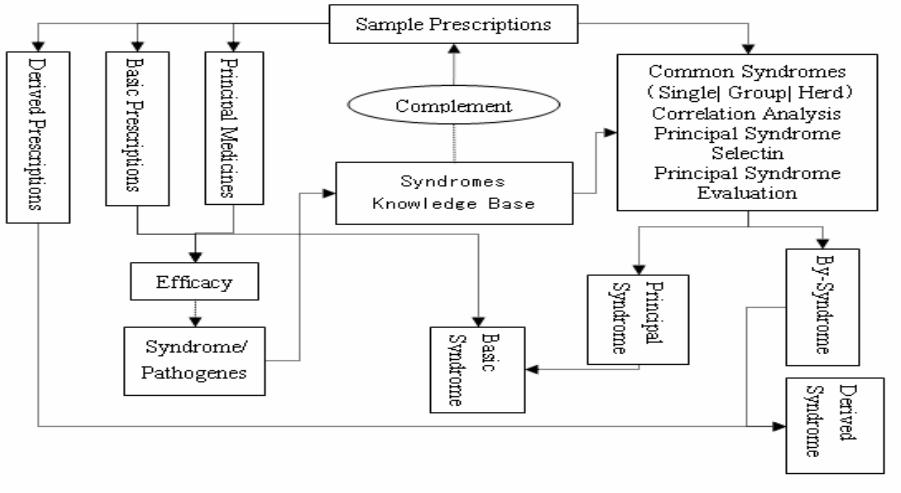

Fig. 4. Technique routine of prescription syndrome correlation analysis

3. Design Issues

The TCM data that need to be analyzed are got through incipient handy input about 2 years ago in Microsoft Access format where 1355 Spleen-stomach prescriptions were screened out from Dictionary of Traditional Chinese Medicine for sample data of analysis. However, most of these data fields are text attributed including some fundamental fields, such as medicine effect, Quality, Flavor, Channel tropism, etc. which make the analyses intangible and unfeasible regarding that most analysis methods are designed for continuous values. In addition, many textual expressions are far from clear and regular. For example, the syndrome item "tiredness" may have many different terms across the whole syndrome database with similar meanings, such as "weary", "exhaustiveness", etc. But these terms are often used independently and have corresponding specific applicable environment respectively. It is impossible to define the concepts of these terms as well as their governing scope which is said to be "perceptible but not expressible", however, to analyze these data in computer fashion, every trivial must be strictly formulated. Even among the rare countable fields, such as Quality, Flavor, Channel tropism, etc. which could be easily converted into categorical numbers, such expression still lag behind what the domain professional really want out of them because the categorical representations did not cover all their implications. Moreover, the TCM data sampled by the task is only a small portion of the huge amount of TCM data. What the system needs to face is large scale data bases containing thousands and thousands of prescriptions. So the designing and implementing of highly efficient and scalable algorithms remains the core of the whole project at least from the angle of the development process of this system though some may accentuate the system's righteousness while ignore its performance. At last, to best satisfy the users' requirements, the system must be designed highly flexible and have good interoperability. Based on the abovementioned considerations, the following issues have to be addressed:

3.1 Data standardization and regularity

There are a lot of ambiguities and irregularities in the original prescriptions data, mainly comprising the following two kind of occurrences: (1) Semantic ambiguities, i.e. some traditional definitions, terms or expressions need to be reanalyzed and positioned so as to take on a clearer and more sensible meaning. (2) Expression irregularity, i.e. many different expressions can be used to imply or indicate the same definition (namely, multiple terms one concept). To solve this problem, we adopt the following scheme to regulate and integrate the source data. For the one-term-multiple-concepts case, we eliminate the term from alphabet and reserve only the expressions for detailed concepts. And as to the multiple-terms-one-concept case, we only reserve the definition for the uniformed concept. By doing this, we unify the concepts' space and avoid the consequences caused by the inconsistency of the original data.

3.2 Handling categorical data

Most data analysis algorithms are designed to run on continuous data, however a large portion of the source TCM data is categorical. It then becomes a problem to measure the value of these categorical items. The key issue is to develop a scientific and practical data assessment scheme or system which transforms the categorical data items into continuous values. In the development of TCMiner, the establishment of such a system can be divided into two steps: firstly, try to preset the upper and lower bound of target continuous value scope and its corresponding mapping function with according to given constrainsts of the TCM knowledge; secondly, suppose the above digitalization scheme to be true and train the scheme by conducting sample test on prepared testing data; finally, test the trained scheme by validating on prepared testing data. The training and testing method can be justified by the introduction of highly precise and fast GEP approaches.

3.3 Adoption of high performance algorithms

Traditional Chinese Medicine has a treatment history of more than 3000 years and there have accumulated thousands of thousands of ancient prescriptions spanning across hundreds of TCM works. On the other hand, most valuable patterns are obtained through deeply and sophisticated analyses involving different aspects (Prescriptions, Medicines, Syndromes, etc.). To mine complicated knowledge through such a huge amount of data, the algorithms adopted must have high processing performance and scalability as well.

3.4 Flexibility and Interoperability considerations

The very significant requirements for a data mining system are to provide enough flexibility and scalability. When we construct the system, one of the key points in our minds is to make it highly generalized so that it can be flexible enough to adapt itself to the diverse needs of users. Another key point is how to provide multiple data analysis technologies, i.e. mining engines so that users can select the appropriate one according to the characteristics of the data to be analyzed and the way in which the extracted information will be used for different TCM data analysis tasks involve different datasets and different datasets may have different formats in

case of wide existence of legacy database, text dataset, excel forms established at different times.

4. System Features

In the development of TCMiner,we use the latest object Oriented techniques to achieve a high degree of portability, accessibility and maintainability. The implementation in Delphi allows the system to have high developing speed and run on multiple operating systems including Microsoft Windows 98, 2000 and XP Linux, and UNIX.

4.1 Process Flow

The TCMiner system architecture can be viewed as processes on a data stream in that the mining tasks are broken down into a series of subtasks with results from each step passed to the next. Since existing TCM prescriptions may be save in different databases and have different formats and usually there are noisy data and inconsistencies, the first step toward knowledge is to clean and integrate the data. In this process, TCMiner finishes the following tasks: filling out vacant values, smoothing away noise data, removing the inconsistency and reducing the task irrelevant data. The second step is to construct TCM data warehouse, which is of great essence to the multi-dimensional data analysis. In TCMiner, four data marts are enclosed in the TCM data warehouse and their subjects are the prescription, medicine, syndrome and Pathogenesis respectively. Due to limited space, we only give the star scheme of Prescription Mart shown below.

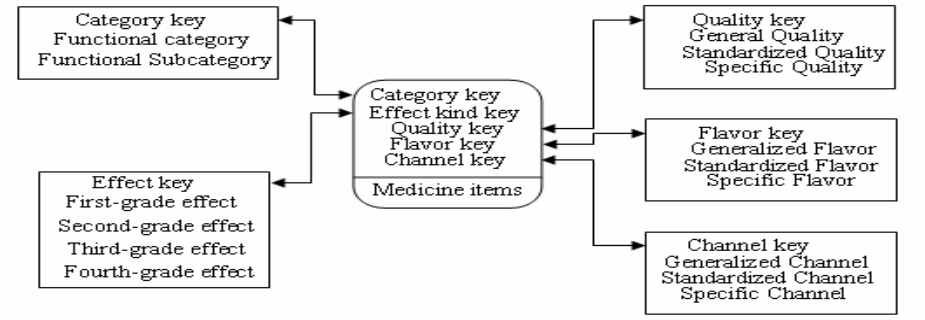

Fig. 5. Star scheme of Prescription Data Mart

The third step is execution of the essential tasks performed by the Multi-Dimension Analysis Engine or the OLAP Engine mainly on the TCM data warehouses. Finally, the patterns are presented to the Patterns Evaluation Module to make the last selection. And only the rules passing the filtering would have the opportunity to show on the end user's screen.

4.2 Major Components

The Multi-Dimensional Data Analysis Engine is responsible for most multi-dimensional data analysis tasks. For example, it can accomplish requests such as "What medicines and herbs are the building blocks of a typical spleen-stomach prescription treating menopause with the Quality to be cold and Flavor to be light bitter?" The Multi-dimensional Data Analysis Engine

interprets a mining task specification which provides the details about each specified operation and their executing sequence and put to practice.

The OLAP Engine is responsible for providing a simple, summarized and generalized view of data as a whole through rolling up, drilling down, slicing, dicing, and pivoting operations according to user-specified conditions. The OLAP Engine works in corporation with the Multi-dimensional Data Analysis Engine at times when the tasks involve processing on both aspects.

Validation Tester works as a safeguard to check the user's request to see if there is grammatical mistakes, internal contradictions, and semantic errors or if the request is out of the system's processing capacity. If one of the abovementioned things happens, the request will be regarded as invalid and returned to the user with detailed situation description. Otherwise, the request will be added to the tasks queue of Task Scheduler with specific description of resource requirements. And the Task Scheduler will do the subsequent post-processing.

Task scheduler is in charge of maintaining and scheduling management of the task list. It examines task list and determines which task should be executed. The scheduling strategy can be customized to adapt to different system requirements.

Knowledge base works as a heuristic guide for TCM knowledge mining. The knowledge base embedded a considerable set of TCM knowledge rules containing the correlations of the prescriptions with certain Quality, Flavor, Channel tropism and their corresponding medicines and their respective syndromes in form of rules.

Patterns Evaluator reexamines the resulting patterns in the following four aspects: (1) availability: the resulting patterns should be validated according to corresponding rules of the Knowledge Base (2) size: association mining usually generates too many rules which makes it difficult for the users to pick valuable ones. Pattern Evaluator checks the resulting rules' size to see if it is relatively too large, and returns the result for refinery mining. (3) user-specified template: TCMiner supports user-specified template and extra constraint in associations mining. (4) Grammatical and consistency test: this process finish the formulary examination of the resulting patterns.

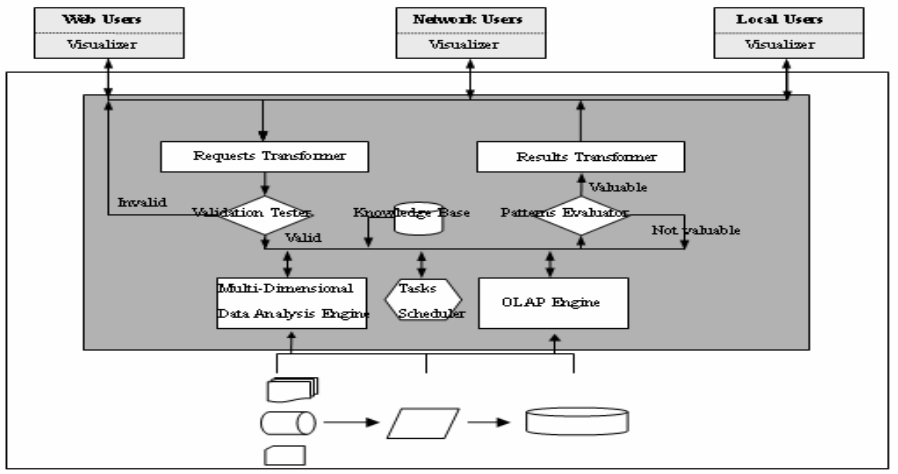

Fig. 6. General architecture of TCMiner

5. Performance Study of Implemented Algorithms

In this section, we present the performance study of the major algorithms implemented in TCMiner over a variety of datasets. All Algorithms are implemented in C++. Readers are recommended to refer to paper [3]-[6] for detailed design thought, pseudo code, example of these algorithms, which are our progresses at the previous stage. Due to limitation of space, we only demonstrate performances of them. All DataSets involved are the standard testing data which can be downloaded through websites [8]-[9] including T10K, T100K, T100K-800K, Connect-4, Pumsb, and Mushroom.

5.1 Frequent Patterns Mining Algorithm

The following charts contrast the performance of our algorithm with the almost best efficient algorithm FP-Growth [10]. The experiments are done on a Windows 98 machine with CPU and memory to be Intel 450M and 128M respectively. As is shown in the first two figures, in both synthetic and real world dataset (dense datasets) our algorithm outperforms FP-Growth greatly and sharply. The subsequent two graphs of experiments on T10K-T100K datasets ensures that our algorithm has better scalability in both time and space perspectives.

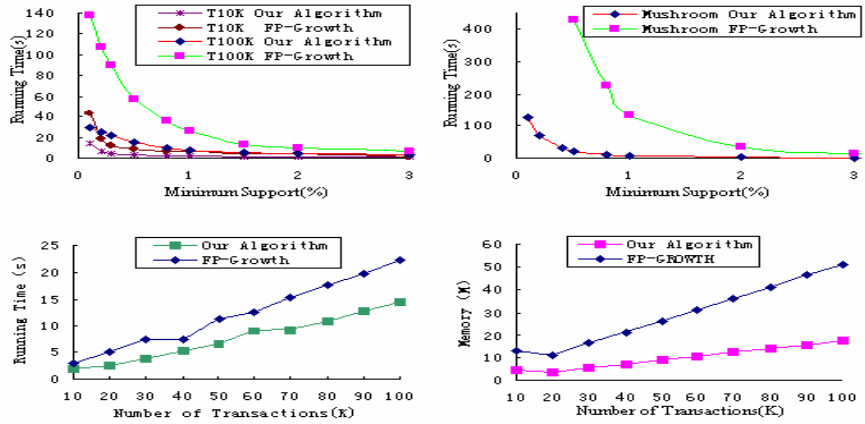

Fig. 7. Performance and scalability contrast of our algorithm and FP-Growth

5.2 Association Rules Generation Algorithm

The following experiments show the performance of our algorithm—SPF in comparison with the latest version Apriori [5] algorithm in both speed and scalability aspects. The experimental environment is: OS Windows 2000, CPU Intel 900M，Memory 128M. This group of experiments shows that our algorithm runs far faster than Apriori algorithm and has far better excellent scalability.

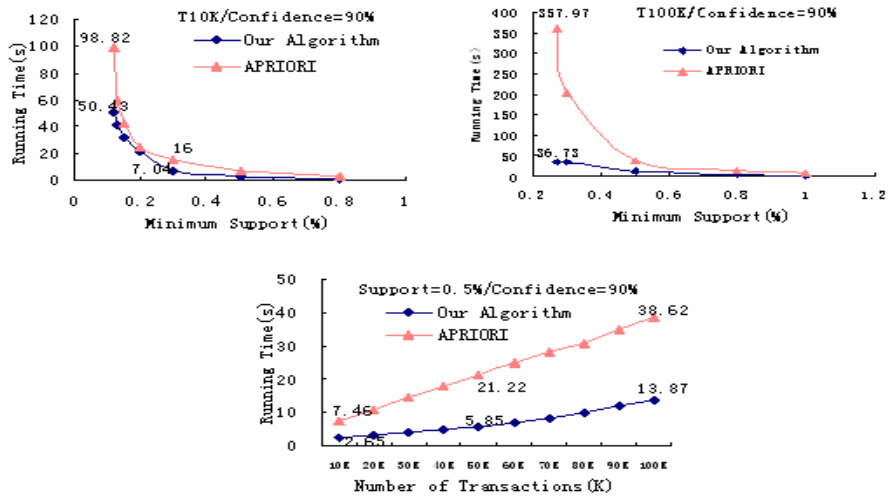

Fig. 8. Performance and scalability contrast of our algorithm and latest APRIORI

5.3 Frequent Closed Patterns and rules Mining Algorithm

In paper [6], we propose an efficient algorithm for frequent closed itemsets mining. In addition, we have extended FCIS into CI_RULES, which can highly efficiently produce frequent closed itemsets rules. The following experiments were done on a 233MHz PC with 128MB of memory, running Windows 2000 to compare both performance and scalability of our algorithm and CLOSET on Pumsb and Connect-4. Please note that the time axes are all in logarithmic scale. The performance data of CLOSET and CHARM is extracted from paper [12].

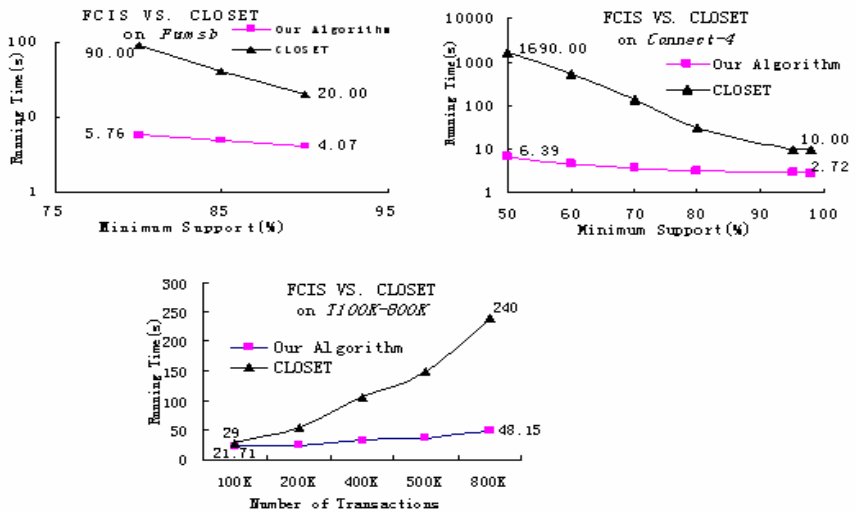

Fig. 9. Performance and scalability contrast of our algorithm and CLOSET

These experiments show: (1) our algorithm outperforms CLOSET with at least an order of magnitude on real datasets; (2) our algorithm has a wonderful scalability; (3) our algorithm can highly efficiently produce frequent closed itemsets rules.

6. Future Directions

Although TCMiner has achieved inspiring results in both computer and TCM sides, some problems still exist. One noticeable problem is that for most theoretical prescription research the clinical application oriented issues are always ignored to some extent, which leads to the separation of research of TCM prescription paring laws and the actual TCM clinical practice and hence affects the value and quality of TCM paring research. The only way out is to integrate more information on the clinical practice, esp. the particular responses of the patients, build proper describing models, and put more emphasis on individual patient group reactions. In addition, the adoption of clinical and experimental methods to discover the single medicine and small amount prescription paring rules is the major fallback in the medicine trade up to now. Our near future work will try to ameliorate TCMiner into a more sophisticated TCM mining system with more comprehensive functions.

7. Conclusion

This paper introduces the architecture and algorithms of TCMiner— a high-performance data mining system for the multi-dimensional data analysis of Traditional Chinese Medicine prescriptions developed by the Data Base and Knowledge Engineering Lab (DBKE) at Computer School of Sichuan University. The system has the following features: (1) High Performance (2) Multi-dimensional Analysis (3) Flexibility (4) Interoperability (5) Optimization for TCM. This data mining system can work as a powerful assistant for TCM experts by conducting Traditional Chinese Medicine Data Mining such as Computer Aided Medicine Pairing Analysis, Medicine Syndrome Correlation, Quality and Flavor Trend Analysis, and Principal Components Analysis and Prescriptions Reduction etc.

References

1. General Guidelines for Methodologies on Research and Evaluation of Traditional Medicine, http://www.who.int/medicines/library/trm/who-edm-trm-2000-1/who-edm-trm-2000-1.pdf
2. Guste Editors' Notes on the special issue, http://www.sinica.edu.tw/~jds/preface.pdf
3. Fan Ming, Li Chuan, Mining Frequent Patterns in an FP-tree Without Conditional FP-tree Generation, Journal of Computer Research and Development, 40th Vol. 2004
4. Cai guoqiang, Li Chuan, Fan Ming, A NEW ALGORITHM ON MULTI-DIMENSIONAL ASSOCIATION RULES MINING, Journal of Computer Science, Aug. 29th Vol. A Complement, page 1-4, 2002
5. Li Chuan, Fan Ming, GENERATING ASSOCIATION RULES BASED ON THREADED FREQUENT PATTERN TREE, Journal of Computer Engineering and Application, 4th Vol., 2004
6. Fan Ming, Li Chuan, A Fast Algorithm for Mining Frequent Closed ItemSets, submitted to ICDM'04
7. Li Chuan, Fan Ming Research on Single-dimensional Association Mining, Full Paper Data Base of Wanfang Network
8. http://www.ics.uci.edu/~mlearn/MLRepository.html

9. http://www.almaden.ibm.com/cs/quest/demos.html
10. J. Han, J. Pei and Y. Yin. Mining frequent patterns without candidate generation. Proc. 2000 ACM-SIGMOD Intl. Conf. on Management of Data, pages 1-12. May 2000.
11. R. Agrawal and R. Srikant. Fast algorithms for Mining association rules. Proc. 1994 Int'l Conf. on Very Large Data Bases, pages 487-499, Sept. 1994.
12. Jian Pei, Jiawei Han, and Runying Mao. CLOSET：An Efficient Algorithm for Mining Frequent Closed Itemsets. Proc. 2000 ACM-SIGMOD Int. 2000 ACM SIGMOD Intl. Conference on Management of Data. page 8-10