Discovering Multi-dimensional Major Medicines from Traditional Chinese Medicine Prescriptions *

LI Chuan[1], TANG Changjie[1], ZENG Chunqiu[1], WU Jiang[1], CHEN Yu[1], QIU Jiangtao[1], ZHU Jun[2], DAI Li[2], JIANG Yongguang[3]
[1]*The Database and Knowledge Engineering Lab, Computer School of Sichuan University*
[2]*National Centre for Birth Defects Monitoring, Sichuan University*
[3]*Chengdu University of Traditional Chinese Medicine*
{lichuan, tangchangjie}@cs.scu.edu.cn

Abstract

Multi-dimensional major medicines analysis is one of the most important tasks in the data analysis of Traditional Chinese Medicine (TCM) prescriptions. In this paper, an effective method is proposed to mine multi-dimensional major medicines from TCM prescriptions. The main contributions include: (1) proposing the concept of multi-dimensional major medicines, (2) borrowing the concept of multi-dimensional frequent patterns and improving the approach of Multi-dimensional Index Tree, (3) applying it in the TCM major medicines mining, (4) implementing the algorithms in the major medicines discovery module of TCMiner 1.0. (5) Extensive experiments show the effectiveness and efficiency of the proposed approach.

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]. However, despite its existence and continued use over many centuries, its bio-chemical mechanism and formula synergic effects are still a mystery at least in theoretical sense because of its complicated physiochemical composition [2].

Discovering multi-dimensional major medicines can help TCM experts explore the structure of TCM prescriptions and provide valuable guidance in new prescriptions invention. To better explain the concept of multi-dimensional major medicines, we introduce the idea of multi-dimensional frequent pattern.

Example 1.1. (**Target Marketing**) A supermarket manager may ask a question like "what groups of customers would like to buy what groups of items?" What he wants to know is not as simple as what commodity items tend to be sold together, but (1) what kinds of customer groups are there, and (2) the major interests of each of the different customer groups. So what concern him are the patterns like the following (Suppose the minimum support threshold is 2%).

(1) {*BMEI Proceedings*, *TCMiner version 1.0*} ∧ {*Occupation*('*Student*'), *Major*('*Computer*')}: *2%*

(2) {*Hamburg*, *Milk*, *Egg*, *Meat*, *Edible Oil*} ∧ {*Occupation*('*House Wife*'), *Income*('*Middle*')}: *4%*

……

Each pattern has shown a different customer group and its major shopping interests. These patterns are different from traditional frequent patterns [3] in that they have the following three features.

(1) They contain transaction item sets, such as {*BMEI Proceedings*, *TCMiner version 1.0*} and {*Hamburg*, *Milk*, *Egg*, *Meat*, *Edible Oil*}, etc.

(2) They contain the multi-dimensional values combinations, such as {*Occupation* ('*Student*'), *Major* ('*Computer*')} and {*Occupation* ('*House Wife*'), *Income* ('*Middle*')}, etc.

(3) The number of occurrences of both itemsets and multi-dimensional values combination is above the user-given minimum support count.

Such patterns have clearly represented the major customer groups and the interests of different customer groups. Paper [16] has formally proposed and defined the multi-dimensional frequent patterns. Multi-dimensional major medicines are the major medicines of the different major multi-dimensional prescription groups. This paper uses multi-dimensional frequent pattern to represent the multi-dimensional major medicines.

There are special difficulties to mine multi-dimensional frequent patterns by traditional *Apriori* [6] or *FP-Growth* [3] algorithms.

(1) These algorithms produce too many invalid patterns, such as those containing only itemsets and those containing only dimension values combinations.

* This work was supported by NSFC Grants (60773169 and 90409007), 11th Five Years Key Programs (2006038002003) of China, the Tackle Key Problem Fund (2006Z01-027), and Sichuan University Youth Foundation (06036).

978-0-7695-3118-2/08 $25.00 © 2008 IEEE
DOI 10.1109/BMEI.2008.244

(2) The resulting patterns are not grouped according to dimensions of users' concerns. Since frequent patterns are usually in a large number, such disorder is not acceptable [4, 7, 8].
(3) Technical problems such as (a) if we use *Apriori* algorithm, the system performance will be a bottleneck as *Apriori* needs to scan database multiple times [6].(b) if we use *FP-Growth*, it can not distinguish a pattern among its many possible explanations because in realistic databases, the coding of different dimensions are usually overlapped [3, 4, 8].
(4) There is an underlying hypothesis that the closeness degree between the resultant itemsets and multi-dimension values combinations should be adjustable, which single-dimension algorithms cannot satisfy [7, 8, 9].

Therefore, almost all existing approaches for multi-dimensional frequent patterns mining rely on the construction of data cube [5, 10, 11, 12]. Unfortunately, the storage space of data cube grows explosively as dimensionality or cardinality (number of distinct values in a dimension) grows as shown in Example 1.2.

Example 1.2 Suppose a data cube consists of 6 dimensions, with cardinality of 1000. Then it contains $(1001)^6$ cube cells [4, 5, 13]. Although the adoption of partial materialization such as iceberg cube, condensed cube, etc. can delay the space growth, the fundamental space explosion caused by data cube construction can not be solved [10, 11, 12, 13].

To deal with the problem, this paper borrows the concept of multi-dimensional frequent patterns and the approach of Multi-dimensional Index Tree, applies it in the major medicines mining, and implements the algorithms in the major medicines discovery module of TCMiner 1.0. Extensive experiments show the method is effective and efficient.

2. Multi-dimensional TCM Major Medicine

Designing an efficient algorithm for multi-dimensional major medicines analysis of TCM prescriptions has been considered by both the TCM and data mining trade for quite a long period. 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) of Sichuan University has been working on many projects (SATCM 2004JP40, NSFC Grants 90409007, etc.) in collaboration with Chengdu University of Traditional Chinese Medicine to investigate new methods for multi-dimensional major medicines discovery from TCM prescriptions [14, 15].

Example 2.1 Figure 1 depicts the logical view of a simplified TCM prescription [8].

Name	*Ginseng Defensive Soup*
Ingredients	*Root of herbaceous peony*
	Atractylodes macrocephala
	Astragalus
	Ginseng
	Zhigancao
Functions	*Resolving heat*
	Eliminate dampness
Indications	*Treating malaria*
Usage	*Boiled taken*

Fig. 1. A simplified TCM prescription

This prescription contains *5* medicine items, i.e. *Zhigancao*, *Ginseng*, etc. and *3* overall properties, *Resolving heat* and *Eliminate dampness* (function), *Treating malaria* (indication) and *Boiled taken* (usage). The physical TCM prescription database is a multi-dimensional database comprising about *20* tables in Microsoft Access format with more than *100,000* prescriptions. The conceptual schema is a multi-dimensional structure as shown in Figure 2.

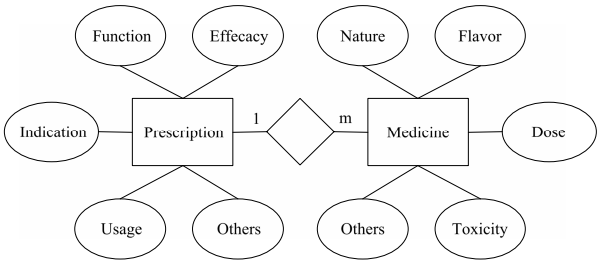

Fig. 2. The conceptual schema of TCM database

The General demand of multi-dimensional TCM major medicines analysis is to discover the major kinds of prescriptions and the major medicines in each kind, which can be translated to the following questions.
(1) What major kinds of prescriptions are there in the TCM prescriptions database (E.g. {*Function* (*'resolving heat'*), *Indication* (*'Treating malaria'*)} is a kind of prescriptions)?
(2) What are the major medicines for each kind of prescriptions (E.g. *Atractylodes macrocephala* and *Astragalus* are major medicines of prescriptions kind {*Function* (*'resolving heat'*), *Indication* (*'Treating malaria'*)})?

In order to systematically answer these questions, this paper proposes the concept of multi-dimensional frequent pattern and studies its mining method.

3. The concept of the Multi-dimensional Frequent Pattern

Paper [16] proposes the concept of multi-dimensional frequent patterns as explained below. Suppose there are D dimensions, denoted as $\{P_1, P_2, ... P_D\}$. Let I be a set of items, p_i be a possible value of the i_{th} dimension. Let $l, l1, l2, \ldots lj$ be distinct integers between 1 and D. We have the following definitions.

Definition 3.1 (Multi-dimensional Pattern) Let i, i_1, $i_2, \ldots i_k$ be items of I. Then $i \wedge \{p_{l1}, p_{l2}, \ldots p_{lj}\}$ is called a multi-dimensional item and $\{i_1, i_2 \ldots i_k\} \wedge \{p_{l1}, p_{l2}, \ldots p_{lj}\}$ is called a multi-dimensional pattern**,** where $\{p_{l1}, p_{l2}, \ldots p_{lj}\}$ is called a multi-dimensional values combination.

Definition 3.2 (The Universal Set) The union of all distinct multi-dimensional patterns of the form $I \wedge \{p_{l1}, p_{l2}, \ldots p_{lj}\}$ is called the universal set of the multi-dimensional space, denoted as MI,

$$MI = \bigcup_{possible\ l1,l2,\ldots lj} \left(I \wedge \{p_{l1}, p_{l2}, \ldots p_{lj}\}\right).$$

Definition 3.3 (Sub-pattern Relationship) Let i_1', $i_2', \ldots i_k'$ be items of I, and $l', l1', l2', \ldots lj'$ be distinct integers between 1 and D. Let $M_A = \{i_1, i_2 \ldots i_j\} \wedge \{p_{l1}, p_{l2}, \ldots p_{lj}\}$, $M_B = \{i_1', i_2' \ldots i_k'\} \wedge \{p_{l1'}, p_{l2'}, \ldots p_{lj'}\}$ be two multi-dimensional patterns. M_A is called to be sub-pattern of M_B, denoted as $M_A \subseteq M_B$, if the following two criteria hold simultaneously:

(1) $\{i_1, i_2 \ldots i_j\} \subseteq \{i_1', i_2' \ldots i_k'\}$, and

(2) for any element $p_l \in \{p_{l1}, p_{l2}, \ldots p_{lj}\}$, there exists an element $p_{l'} \in \{p_{l1'}, p_{l2'}, \ldots p_{lj'}\}$[2] such that (a) p_l and $p_{l'}$ are values of the same dimension and (b) $p_l = p_{l'}$.

Particularly, for the universal set, MI, the sub-pattern relationship is defined below.

A pattern $M_A = \{i_1, i_2 \ldots i_j\} \wedge \{p_{l1'}, p_{l2'}, \ldots p_{lj'}\}$ is called a sub-pattern of MI**,** denoted as $M_A \subseteq MI$, if one of the following two criteria holds.

(1) M_A is of the form $I \wedge \{p_{l1}, p_{l2}, \ldots p_{lj}\}$ or

(2) There exists an multi-dimensional pattern, $MB = I \wedge \{p_{l1}, p_{l2}, \ldots p_{lj}\}$ satisfying the following two criteria simultaneously:
   (a) $M_B \subseteq MI$, and
   (b) $M_A \subseteq M_B$

Based on Definition 3.2 and 3.3 it's clear that for any multi-dimensional pattern M_A**,** $M_A \subseteq MI$**.**

Definition 3.4 (Multi-dimensional Transaction Database) Given the universal set MI, a set of multi-dimensional patterns $\{mt_1, mt_2, \ldots mt_n\}$ is called a multi-dimensional transaction database, denoted as $MTDB$, if for any $i \in [1..n]$, $mt_i \subseteq MI$. The support η (or occurrence frequency) of a multi-dimensional values combination $\{p_{l1}, p_{l2}, \ldots p_{lj}\}$ is the number of transactions containing $\{p_{l1}, p_{l2}, \ldots p_{lj}\}$ in $MTDB$. The support η of a multi-dimensional pattern MA is the number of transactions containing MA in $MTDB$.

Definition 3.5 (Multi-dimensional Frequent Pattern) Let η be the predefined minimum support threshold. A multi-dimensional values combination is frequent if its support is no less than η. A multi-dimensional pattern is a multi-dimensional frequent pattern if its support is no less than η. The problem of multi-dimensional frequent patterns mining is to generate the complete set of multi-dimensional frequent patterns from $MTDB$.

4. The Concept and Algorithms of MDIT

Based on the discussion above, we introduce here the Multi-dimensional Index Tree, and two new algorithms: one for computing MDIT, and the other for multi-dimensional frequent patterns mining based on MDIT.

The general idea is to partition the base dataset into a series of projected databases according to different frequent multi-dimensional values combinations. With MDIT, one can find frequent patterns from each projected database and putting out in conjunction with corresponding multi-dimensional values combination, one can get the complete set of multi-dimensional frequent patterns.

4.1 Multi-dimensional Index Tree (MDIT)

MDIT is a tree structure. Internal nodes of MDIT are composed of several inverted index items. The leaf nodes are transaction buckets.

Definition 4.1 (**Inverted Index Item**) An inverted index item is a 2-tuple $R = \{I, P\}$, where I denotes a dimension value and P is a pointer leading to a child node. The inverted index is represented by a series of inverted index items at a branch of nodes from root to leaf.

Definition 4.2 (**Multi-dimensional Index Tree**, denoted as **MDIT**) MDIT is a tree defined as follows.

(1) MDIT consists of a root, a set of internal nodes, and a set of leaf nodes.

(2) The root and the internal nodes are composed of a series of inverted index items.

(3) All the leaf nodes are transaction buckets.

(4) A transaction bucket is a tuple $B = \{S, TIDs\}$, where S denotes the bucket frequency, and $TIDs$ is a *TID list*.

Definition 4.3 (**MTDB Array**) Suppose there are n transactions $\{mt_1, mt_2, \ldots mt_n\}$ in $MTDB$. A $MTDB$ array is a global array $A = \{t_1, t_2, \ldots t_n\}$, where for each i

[2] Just having $\{p_{l1}, p_{l2}, \ldots p_{lj}\} \subseteq \{p_{l1'}, p_{l2'}, \ldots p_{lj'}\}$ is incorrect because it may not satisfy (a).

$\in [1..n]$, t_i is the itemset part of mt_i. The *MTDB* array can be easily organized into a relational DB table or directly in the memory.

4.2 Construction Algorithm of MDIT

Based on above illustration, the MDIT building algorithm can be summarized below.

Algorithm 4.1 Build-MDIT
Input: A multi-dimensional database $MTDB = \{mt_1, mt_2, ..., mt_m\}$ with n dimensions $(P_1, P_2, ..., P_n)$.
Output:
(1) A MDIT and
(2) A *MTDB* array
Method:
(1) Initialize the MDIT with *root* and an empty *MTDB* array;
(2) **for** each $mt_i \in MTDB$ {
(3) **for** each multi-dimensional values combination $\{p_{1l}, p_{2l}, \ldots p_{lj}\}$ contained in mt_i
(4) **if** there is no branch in MDIT representing $\{p_{1l}, p_{2l}, \ldots p_{lj}$: n: *TID list*$\}$[3],
(5) Generate a new branch from root to leaf to represent $\{p_{1l}, p_{2l}, \ldots p_{lj}$:*1*: $mt_i\}$;
(6) **else** {
(7) Register the *TID* of mt_i into *TIDs* of the bucket representing $\{p_{1l}, p_{2l}, \ldots p_{lj}$: n: *TID list*$\}$;
(8) Increase the bucket frequency by 1.
(9) } // **end of else**
(10) Add the *TID* and the itemset part of mt_i into the *MTDB* array;
(11) } // **end of for**

4.3 Algorithm MDIT-Mining

Given the MDIT, one can mine multi-dimensional frequent patterns directly. Just as in case of inverted indices, firstly search the MDIT to find all the frequent buckets. Secondly, for each frequent bucket, get its projected database via its *TID list*, mine the frequent patterns wherefrom and finally output them with their corresponding multi-dimensional values combinations. This leads to our algorithm for multi-dimensional frequent patterns mining as below.

Algorithm 4.2 (MDIT-Mining) Mining multi-dimensional frequent patterns based on MDIT
Input: MDIT and the minimum support η.
Output: The complete set of multi-dimensional frequent patterns.
Method:
(1) **for** each transaction bucket B {
(2) **if** (B->S >= η) { // Consider frequent buckets
(3) *Mdvc* = the multi-dimensional values combination represented by inverted index items at a MDIT branch from root to Bucket B;
(4) Construct *FP-Tree* with the projected database represented by *B->TIDs*;
(5) *FP* = *FP_Growth* (*FP-Tree*, η);
(6) **for** each pattern $p \in FP$
(7) Output $\{p\} \wedge Mdvc$
(8) }}

5. Major Medicine Module in TCMiner 1.0

This section reports the implementation and applications of Build-MDIT and MDIT-Mining in the TCMiner1.0. Due to space limitation, many related synthetic data experiments are omitted here. Any researcher who want to learn details, please feel free to contact me at Charles.li2004@gmail.com.

5.1 Experiments on Real TCM Datasets

The tests are conducted on two real datasets. The first dataset contains *2870* fundamental prescriptions with *6* dimensions (*source, dynasty, form, efficacy, implication, Symptom*). The cardinalities are (*352, 9, 103, 1551, 1207, and 1074*). The MDIT is built in *0.19* seconds with *14.3MB*. And MDIT-Mining took less than *0.89* seconds in outputting *84555* multi-dimensional frequent patterns. The second dataset contains *12037* prescriptions with *4* dimensions. The MDIT took *0.28* seconds and *10.7MB*. In both experiments, the minimum support is *0.1%*. The mining time is *0.27* seconds for mining all *2892* patterns.

5.2 System Implementation

Traditional Chinese Medicine Miner version 1.0 (TCMiner 1.0) is developed using *Delphi 6.0* / *Microsoft Access 6.0* [14]. The system can fulfill many TCM analysis tasks such as major medicines discovery, medicine paring analysis [15], etc. The two algorithms are embedded into the major medicines discovery module of the system through APIs based on C. The interface is shown in Figure 12.

(a) Setting parameters …

[3] $\{p_{1l}, p_{2l}, \ldots p_{lj}\}$ has been contained in the MDIT.

(b) Selecting dimensions …

(c) Mining multi-dimensional major medicines …

(d) The multi-dimensional major medicines

Fig.12. Major medicines discovery of TCMiner

The demo version of the system TCMiner is downloadable from the address http://cs.scu.edu.cn. If there's any network problem, please contact the author directly at lichuan@cs.scu.edu.cn.

6. Conclusion

This paper proposed an efficient method to mine multi-dimensional major medicines from Traditional Chinese Medicine Prescriptions. The contributions include: (1) proposing the concept of multi-dimensional major medicines, (2) borrowing the concept of multi-dimensional frequent patterns and the approach of Multi-dimensional Index Tree, (3) applying it in the major medicines mining, (4) implementing the algorithms in the major medicines discovery module of TCMiner 1.0. (5) Extensive experiments show the method is effective and efficient.

7. 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] Han, J., Pei, J., and Yin, Y. Mining Frequent Patterns without Candidate Generation. SIGMOD, 1-12. 2000.

[4] Data Mining: Concepts and Techniques. Jiawei Han, Micheline Chamber, Morgan Kaufmann, Hardcover, ISBN 1558604898

[5] Kamber, J. Han, and J.Y. Chiang. Metarule-guided mining of multi-dimensional association rules using data cubes. KKD'97.

[6] Rakesh Agrawal, et.al Mining association rules between sets of items in large databases. In Proc. of the ACM SIGMOD 1993

[7] R. Baeza-Yates. et. al Modern Information Retrieval. Addison-Wesley, 1999.

[8] Peng Huairen. The First Volume of Great Formula Dictionary of TCM. People's Medical Publishing House. December 1993

[9] G. Piatetski-Shapiro. Discovery, analysis, and presentation of strong rules. In Knowledge Discovery in Databases, pages 229-248, 1991.

[10] D. Barbara and M. Sullivan. Quasi-cubes: Exploiting approximation in multidimensional databases. SIGMOD Record, 26:12-17, 1997.

[11] S. Agarwal, et. al. On the computation of multidimensional aggregates. In Proc. 22nd VLDB, pages 506--521, Mumbai, Sept. 1996.

[12] Y. Sismanis and N. Roussopoulos. The dwarf data cube eliminates the high dimensionality curse. TR-CS4552, University of Maryland, 2003.

[13] Xiaolei Li, Jiawei Han, and Hector Gonzalez. High-Dimensional OLAP: A Minimal Cubing Approach. In VLDB'04.

[14] LI Chuan, TANG Changjie, et al. TCMiner: A High Performance Data Mining System for Multi-dimensional Data Analysis of Traditional Chinese Medicine Prescriptions. (ER 2004), LNCS. 2004

[15] LI Chuan, TANG Changjie. et. al. NNF: an Effective Approach in Medicine Paring Analysis of Traditional Chinese Medicine Prescriptions DASFAA 2005, LNCS.

[16] LI Chuan, TANG Changjie. et. al. Mining Multi-dimensional Frequent Patterns Without Data Cube Construction, PRICAI 2006, LNAI 4099, pp. 251 – 260, 2006.