Research and application on the Data mining technology in medical information systems

Jinhai Zhang

Marine college of Shandong jiaotong university, Weihai , Shandong,China

Keywords: Data warehousing, data mining, association rules, Apriori algorithm, OLAP (On-line Transaction Processing)

Abstract. After several years of vigorous development, theory and application of data mining technology has produced fruitful results. As an application of strong discipline, data mining technologies have permeated all areas of the national economy, caused great attention from academia and industry, have been widely used for different sectors, provide valuable information for decision-making. China data mining research is still in its infancy stage, the application is only focused on some major industries (such as telecommunications, insurance, banking, etc). In the data mining of medical information, more study abroad and domestic, physician data mining to support not only abroad, but also widely used in patients and the hospital's management. Therefore, the study of data mining in medical industry in China has important practical applications.

Introduction

With the application of computer technology in various industries and fields, basic realization of the informatization of the medical industry. Many hospitals have established their own information systems (Hospital Information System, HIS). Hospital information system data over time into a geometric growth in multiples. Those figures covered not only the text but also in medical activities are contained images, sound and other multimedia data, this gives hospitals the information kept in the database expands. Although these data for hospital management, disease diagnosis and research has a very high value. But most of the hospitals in the area of data processing is limited to normal entry, the most basic operations such as delete, search, statistics, and various business modules to produce data for only a single daily data operations, which makes data resources could not be fully utilized.

However is not just in the hospital database system containing patient's physiological and medical information such as medical imaging, but also contains many of the patient information, some more detailed information such as financial information, which makes hospital database contains a great deal of medical value and economic value. Therefore, raise the level of use of these information resources, for the future development of hospital management and medical treatment to provide comprehensive, accurate and timely information for decision-making has become an urgent problem to be solved.

In this context, the health information system database as the main data source, by building multidimensional cubes of data warehouse based on medical information systems and data, data for health information system is multi-level and multi-angle comprehensive analysis, with a view to the utilization of hospital resources reached a new height. Data warehouse from data stored in various places in the source data, load data and cleansing data. Data mining by using data that is stored in the data warehouse to complete a variety of excavation work and integrated analysis to identify medical information or medical management, and the results show in the form of intuitive to end users, for medical staff work to provide timely, accurate and valuable information.

Medical information system data mining methods and technologies

Data mining (data mining,DM), also known as knowledge discovery in databases (knowledge discovery in database,KDD), is a large, incomplete, noisy, vague and random data to extract hidden where, people do not know in advance, but is potentially useful information and knowledge of non-trivial process (Piatetsky-Shapiro 1996). At present, there are a lot of similar terminology and data mining, knowledge extraction, such as data analysis, pattern analysis of archaeological data (data Archeology), data dredging, business intelligence and decision support.

Data mining by can found of knowledge main including: reflect similar things common nature of general type knowledge, and reflect things the aspects features of features knowledge, and reflect different things Zhijian property difference of differences type knowledge, and reflect things Zhijian rely on or associated of associated type knowledge, and according to history and current of data speculated that future data of forecast type knowledge, and reveals things deviated from general exception phenomenon of deviated from type knowledge,. All of this knowledge on the different conceptual levels was found, as the promotion concept tree, from micro to meso-to macro-to meet the needs of different users, different levels of decision-making. Find tools and methods, common classification, clustering, correlation, pattern recognition, visualization, decision trees, genetic algorithms, uncertainty, and so on.

Data warehouse technology based on logical thinking and rigorous mathematics and statistics "acts of scientific judgement and effective" tool. Data warehousing is a "data integration, knowledge management," an effective means. Data warehouses are subject-oriented, integrated, time-variant, and nonvolatile data sets to support management decision-making process. Different from the traditional application-oriented database, the data in the data warehouse are subject-oriented.

There are various forms of data in the data warehouse, it focuses on several common forms of data in the data warehouse: simple stack structure, Web-integrated structures, simplifying the structure of direct and continuous structure.

1. simple stack structure: it is most commonly used in data warehousing, the most simple form of data. It extracted from the application-oriented daily data in the database, and then follow the corresponding topics, set records in the data warehouse.

2. Web-integrated structure: Web-integrated structure, data storage unit is divided into days, weeks, months, years, and several levels.

3. simple structure: its simple stacked files like some, but not a day when integrated into a data warehouse, but the interval, such as one week or one month. It can also be considered a sampling of the database at certain intervals.

4. continuous: the simple structure of two or more consecutive data file organization, another organization continuous data file can be generated.

OLAP is a process of data warehouse based on online analytical processing, is the interface between the user and the data warehouse. OLAP systems can be integrated in different sectors, it is subject oriented, its basic features are: raw data from most basic operational data in the information system; a reasonable response time due to OLAP users mainly managers and business decision makers, only less than the number of users; operation does not depend on the index in the database.

Online transaction processing OLTP and OLAP online analytical processing are currently the two most popular methods of processing the data. General database used primarily OLTP technology, because it is mainly aimed at the treatment of some common services. OLAP technology is mainly used in a data warehouse, this is mainly because the data in the data warehouse is historical, and general analysis in the data warehouse is focused on decision of complex operations and, finally, to present results to the user easy to understand way.

Analysis and design of hospital information data warehouse

As health care costs statistics occupy an increasingly important position in hospital management, how to formulate scientific and reasonable medical costs has become a hospital work one of the emphases and difficulties. Each month in the traditional way takes a lot of effort making medical cost analysis report and use the report to analyze the medical expenses for the month, due to the traditional manual unable to synthesize the data for multidimensional business analysis, it is difficult to drill to flexible medical costs, and thus unable to timely analysis of abnormal reason, affecting the efficiency of hospital.

In view of this situation, in hospital information systems (HIS) based on data mining technique to establish a medical analysis systems, for the costs of treating patients for a particular analysis, further analysis and prediction, understand why anomalies arise, guiding the medical code of conduct.

Paper research of is based on ORACLE of HIS system, in ORACLE database in the developers can established different type of user, for without business needs and user function of different for its distribution corresponding of table, front-end user is through all table distribution different of user role to operation database, this on ensure has database system in the data of rationality and and on data operation of security, also let database of maintenance and the management became more convenient. Set up for different users of the system and its corresponding data tables, namely: public dictionary management, medical records, hospital outpatient management, patient management, orders management, check management, test management, drug control, clinic fees, hospital charges, charge accounts, medical statistics, operations management, health care management, economics and statistics, and so on.

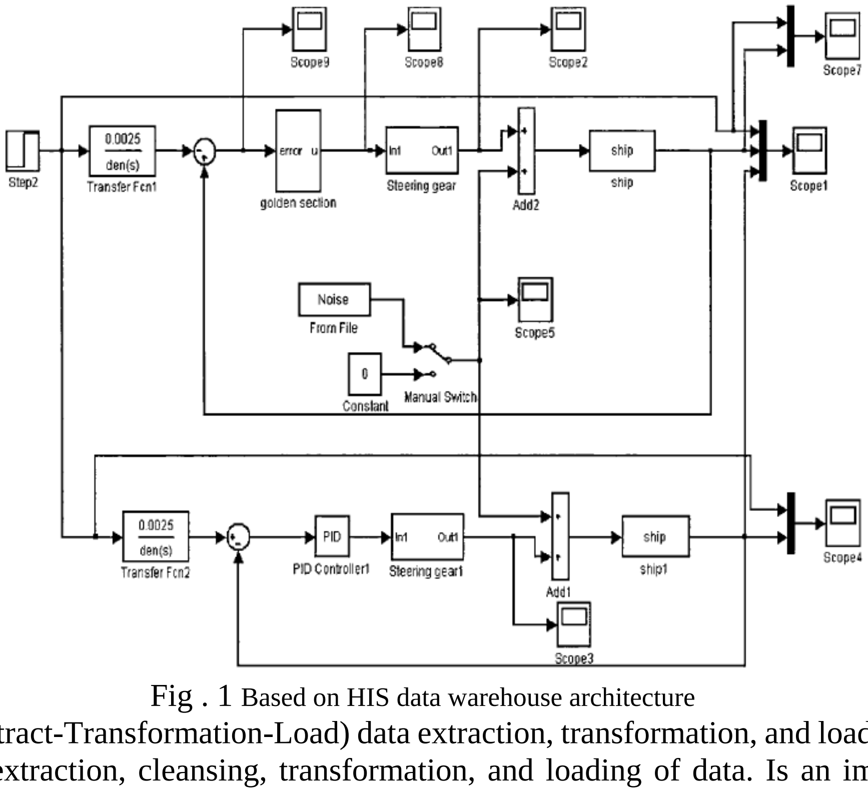


Fig . 1 Based on HIS data warehouse architecture

ETL (ETL,Extract-Transformation-Load) data extraction, transformation, and loading, sometimes called the ETL extraction, cleansing, transformation, and loading of data. Is an important link in building data warehouse. ETL is responsible for the distribution of data from heterogeneous data sources such as relational data, print data after the file is extracted to a temporary middle tier, such as cleaning, transformation, integration, and is loaded into the data warehouse.

Application of data mining in HIS

With the development of HIS system, people began to focus on the accumulated data, from huge data to dig out valuable information through information collection, analysis, and business relationships with business to find the hospital, which provides a comprehensive decision support to hospital management staff, which is the future direction of HIS system.

Is one of the important contents in the data mining of association rules. Paper will associated rules mining application Yu hospital information system among, from large of medical data in the find the level Zhijian of associated relationship, through on gallstones patients data of mining and the analysis, tries to found patients age and medical costs Zhijian of associated relationship, and hope through this associated relationship can let hospital managers more reasonable of using hospital resources, better of control different age level of patients costs, makes hospital benefits maximize.

Association rule mining algorithm considering there are two major problems:

1) reducing the number of I/O operations. Association rule mining algorithm in the number sometimes reached very high levels, too many I/O operations will certainly impact on the efficiency, and to reduce the number of operations, the most important way is to reduce the number of scans the database.

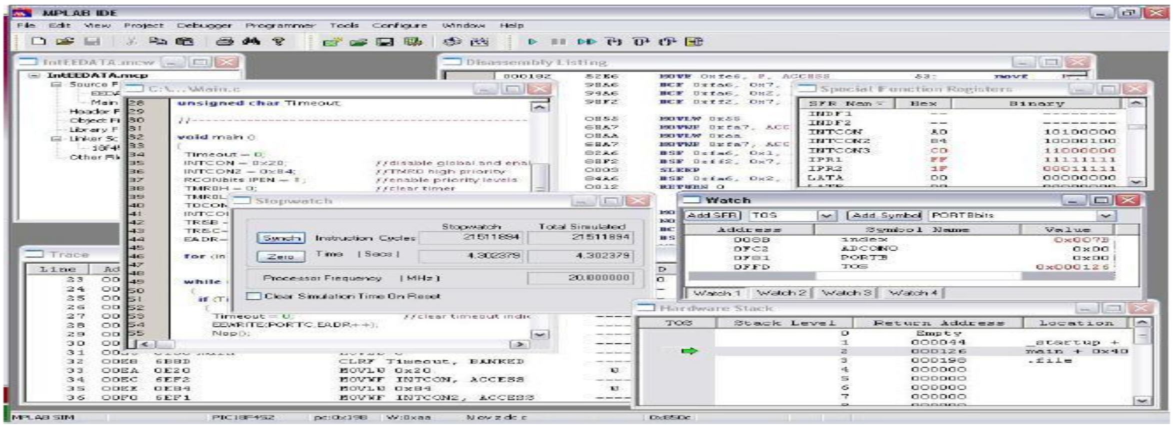
2) candidate set the number of items is too large, the ideal State is broadly similar to their number and frequent itemsets. By reducing the number of candidate itemsets can effectively reduce the set time and storage space.

In 1994, the Agrawal presents Apriori algorithm for discovering frequent itemsets in a database, and its difference with the traditional algorithms mining using a priori knowledge. Agrawal found frequent item sets have 2 very important properties, namely anti-monotonicity properties:

- Frequency must be a subset of the set of frequent item sets.
- Non-frequent item sets a superset of the must is frequent.

Classic shortcomings of Apriori algorithm is: you must take a lot of time on the scan of the database, and decide whether to set the time of accession, are repeatedly to scan the database and compare it with one by one. For small database is acceptable, but when it comes to the amount of data is a huge database, it would have to spend in I/O operations, spending a great deal of growth, even multiply increased. And may allow candidates to become huge.

By optimized algorithms, not only reduces the number of scans the database and reduced generating frequent itemsets candidate item sets, greatly improve the efficiency of the algorithm. he main window as shown in Figure 2:



```
unsigned char Timeout;

//----------------------------------------------------

void main ()
{
  Timeout = 0;
  INTCON = 0x20;                  //disable global and enal
  INTCON2 = 0x84;                 //TMR0 high priority
  RCONbits.IPEN = 1;              //enable priority levels
  TMR0H = 0;                      //clear timer
  TMR0L
  T0CON
  INTCO
  TRISB =
  TRISC =
  EADR =

  for (in

  while
  {
    if (Ti
    {
      Timeout = 0;                //clear timeout indic
      EEWRITE(PORTC,EADR++);
      Nop();
```

	Stopwatch	Total Simulated
Instruction Cycles	21511894	21511894
Time (Secs)	4.302379	4.302379
Processor Frequency (MHz)		20.000000

□ Clear Simulation Time On Reset

```
000182  52E6   MOVF   0xfe6, F, ACCESS        53:   movf
        98A6   BCF    0xfa6, 0x7,
        94A6   BCF    0xfa6, 0x2,
        98F2   BCF    0xff2, 0x7,

        0E55   MOVLW  0x55
        6EA7   MOVWF  0xfa7, ACC
        0EAA   MOVLW  0xaa
        6EA7   MOVWF  0xfa7, ACC
        82A6   BSF    0xfa6, 0x1,
        8EF2   BSF    0xff2, 0x7,
        0003   SLEEP
        84A6   BSF    0xfa6, 0x2,
        0012   RETURN 0
```

SFR Name	Hex	Binary
INDF1	--	--------
INDF2	--	--------
INTCON	A0	10100000
INTCON2	84	10000100
INTCON3	C0	11000000
IPR1	FF	11111111
IPR2	1F	00011111
LATA	00	00000000

Address	Symbol Name	Value
008B	index	0x007B
0FC2	ADCON0	0x00
0F81	PORTB	0x00
0FFD	TOS	0x000126

TOS	Stack Level	Return Address	Location
	0	Empty	
	1	000044	_startup +
➡	2	000126	main + 0x40
	3	000198	.file
	4	000000	
	5	000000	
	6	000000	
	7	000000	

```
Line  Ad
 23   00
 24   00
 25   00
 26   00
 27   00
 28   00
 29   00
 30   00
 31   00
 32   00E8   6BBD          CLRF   Timeout, BANKED
 33   00EA   0E20          MOVLW  0x20
 34   00EC   6EF2          MOVWF  INTCON, ACCESS
 35   00EE   0E84          MOVLW  0x84
 36   00F0   6EF1          MOVWF  INTCON2, ACCESS
```

MPLAB SIM   PIC18F452   pc:0x198   W:0xaa   N ov z dc c   0x850c

Fig . 2 Main interface of the System

References

[1]Chen Wenwei. data warehousing and data mining tutorial [m]. Tsinghua University Press, 2006

[2]Lin Jiebin. theory and practice of data mining OLAP [m]. Tsinghua University Press, 2003

[3]John Shumate. Apraetical Guide to Microsoft 0LAP Server[M]. Addisonwesley. 2000

[4]Gu Yan. based on data warehouse technology in hospital information systems (HIS) Scheme [j]. computer systems applications. 2005

[5]Yi Jing. and implementation of data mining technology in the hospital information [j]. Chongqing University of medical sciences 2007.

[6]Kantardzic M. Data Mining Concept、Models、Methods and Algorithms[M]. IEEE Press.2002

[7] Bryan Ford，Pyda Srisuresh，and Dan Kegel．Peer-to-Peer Communication Across Network Address Translators．In USENIX Annual Technical Conference，Anaheim，CA，April，2005.

[8] Rosenberg J. Schulzrinne H.Camarillo G.Johnston.A.Peterson ， J.Sparks ， and E.Schooler．SIP:Session Initiation Protocol．RFC 3261，June 2002.

[9] Hill R,Wang J,Nahrstedt K.Quantifying Non-functional Requirements:A Process Oriented Approach[C]． IEEE International Requirements Engineering Conference， Kyoto， Japan， 2004：352-353