PAPER

Integration of learning methods, medical literature and expert inspection in medical data mining

Tu Bao HO[†a], , **Saori KAWASAKI**[†b], , **Katsuhiko TAKABAYASHI**[††c], *and* **Canh Hao NGUYEN**[†d],

SUMMARY From lessons learned in medical data mining projects we show that integration of advanced computation techniques and human inspection is indispensable in medical data mining. We proposed an integrated approach that merges data mining and text mining methods plus visualization support for expert evaluation. We also appropriately developed temporal abstraction and text mining methods to exploit the collected data. Furthemore, our visual discovery system D2MS allowed to actively and effectively working with physicians. Significant findings in hepatitis study were obtained by the integrated approach.

key words: *Temporal patterns, temporal relations, visualization, expert inspection, integration, hepatitis study.*

1. Introduction

Medicine has been a traditional domain for artificial intelligence (AI) research and application. It can be observed that the focus on expert systems (ES) in medicine in early days of AI has been changed to intelligent data analysis (IDA) in medicine, especially by machine learning and data mining techniques [5], [16], [22]. At least, two reasons for the new trend are the bottleneck of knowledge acquisition and the explosive growth of medical databases. Intelligent data analysis in medicine has its own features because of the characteristics of medical data. These characteristics include the incompleteness (missing values), incorrectness (noise in data), sparseness (few and/or non-representable patient records available), and inexactness (inappropriate selection of parameters for a given task). Moreover, medical databases are characterized by the particular constraints and difficulties of the privacy-sensitive, heterogeneous, but voluminous, data of medicine [6].

For the last several years we have been working on mining several precious medical databases, especially those on stomach cancer (collected during 1961-1991) at National Cancer Center in Tokyo, and hepatitis (collected during 1982-2001) at Chiba University Hospital. In the first stages, we have developed new computational methods to exploit these databases. The results were promising but always far from the expectation of physicians. The common problem is that most findings were not new to the physicians while what is new to them was usually hard to assess.

Recently, the area of text mining has a considerable progress due to advanced learning methods for complexly structured data. Many works focused on using text mining techniques to exploit voluminous and precious medical literature resources, in particular MEDLINE [7].

Visual data mining [8] is another area that greatly contributes to medical information systems [4]. Appropriate visualization can support deeper understanding of patterns discovered in data, or to uncover their hidden relations [24].

Lessons learned from medical data mining projects [14], [15] have suggested us an integrated approach to medical data mining. The key guideline is that we usually cannot be successful in medical data mining if mining only medical databases with computation methods but multi resources exploitation plus expert participation is indispensable. In fact we have appropriately integrated data mining, text mining, visualization, and expert inspection to do medical data mining.

Section 2 of the paper describes the integrated approach including lessons learned from our medical data mining projects, the integrated approach idea and visual data mining system D2MS. Section 3 presents our study in hepatitis with the proposed integrated approach. The essence is interesting findings due to the integration of multi resources.

2. The Integrated Approach

2.1 Lessons Learned in Medical Data Mining

We have had many unsuccessful attempts in early days of our projects [14], [20]. However, we have gained better insights into these domains from the failures. The following is what we deeply learned in medical data mining: *"It took over twenty or thirty years to collect the medical data, months to understand the problems and preprocess the data, minutes or hours to run data mining, and days or weeks to evaluate the findings"*.

This relation between time of data collection, data mining, and knowledge evaluation suggests data miners to redistribute their effort on these main components

Manuscript received January 18, 2007.
Manuscript revised January 28, 2007.
Final manuscript received January 28, 2007.

a) E-mail: bao@jaist.ac.jp
b) E-mail: skawasa@jaist.ac.jp
c) E-mail: takaba@ho.chiba-u.ac.jp
d) E-mail: canhhao@jaist.ac.jp

of the whole process to obtaining useful medical knowledge. Moreover, we have the following remarks and observations in medical data mining [15]:

- Understanding the domain is crucial and usually requires data miners much effort, time, and collaboration.
- Medical data were usually collected over many years, and lots of them are redundant. Feature selection is usually indispensable, but using only computation does not work well.
- Most knowledge discovered by general-purpose data mining software is not new to the medical doctors.
- There are too many outputs, and model selection indispensably requires users participation.
- The lack of adequate methods and tools limits data mining results.

Our experience shows that there is not always appropriate data mining methods for medical data analysis. From the above remarks and further investigation, we sum up several lessons learned (see [14] for detail):

1. *The knowledge exchange in the collaboration between computer scientists and medical experts is going from tacit to explicit. Such a conversion requires places, occasions, and creations.*
2. *To start with general-purpose data mining software but understand that we may need to create appropriate methods and tools.*
3. *To produce and export learned results in some forms familiar to physicians.*
4. *For the success of medical data analysis, physicians are supposed to participate as much as possible.*
5. *Synergistic visualization of data and knowledge in the data mining process is particularly helpful.*

The lessons are concerned with the central problem of model selection in data mining, i.e., of choosing appropriate discovered models or algorithms and their settings for obtaining such models. In these lessons, we especially recognized the role of physicians in the mining process as well as the crucial importance of background/domain knowledge.

2.2 The Integrated Approach

We consider that in the medical data mining process, the following three factors could be fulfilled in a harmonious combination:

1. Appropriate data mining methods [5], [6];
2. Active participation of physicians supported by a user-friendly environment with visualization tools [3], [8];
3. Effective exploitation of domain knowledge from medical literature [9].

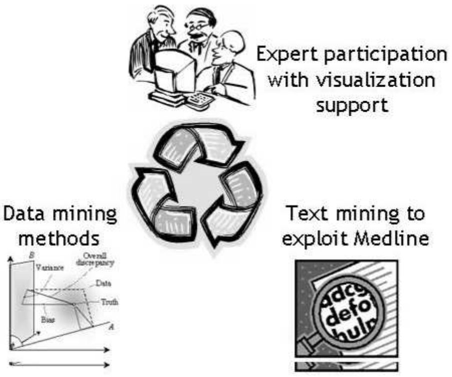

Expert participation with visualization support
Data mining methods
Text mining to exploit Medline
Variance
Overall discrepancy
Data
Truth
Bias
B
A

Fig. 1 Three factors of the integrated approach.

As mentioned in the introduction about the characteristics of medical data, general data mining tools are not always appropriately applied. In many cases, we need either to develop new methods/programs that can work well with the particular data or to select right available tools based on a deep understanding of their power and limit to the data.

Having appropriate data mining tools, the search for new and useful patterns/models from medical data is still a big challenge that usually cannot be well done without an active participation of physicians. It is known but worth noting that a huge possible patterns/models can be induced but they are rarely considered to be novel and useful by physicians. From our best experience, in such a situation the followings are crucial:

- Participation of the physicians in providing guidance or directions such as selected features, heuristics, etc. to narrow the search space when running data mining programs and in evaluating the obtained patterns/models.
- Ease of the expert evaluation on patterns/models discovered. This can be obtained by suitable exported form of discovered patterns/models such as compact tables or visual tools.

The medical literature is one of the sources of background and domain knowledge that we need to exploit. In particular, MEDLINE – the source of life sciences and biomedical information with nearly eleven million records – is very important and can be exploited with text mining methods. Figure 8 illustrates the key idea of this integrated approach.

2.3 A Visual Data Mining System

We have developed the visual data mining system D2MS (Data Mining with Model Selection) as basis for the integrated approach. The key idea of D2MS is to support the user with his/her central role in the data mining process. Essentially, such a system should have support as much as possible to the user in model

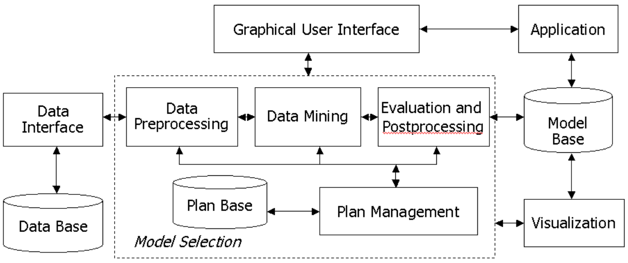


Fig. 2 Conceptual architecture of D2MS

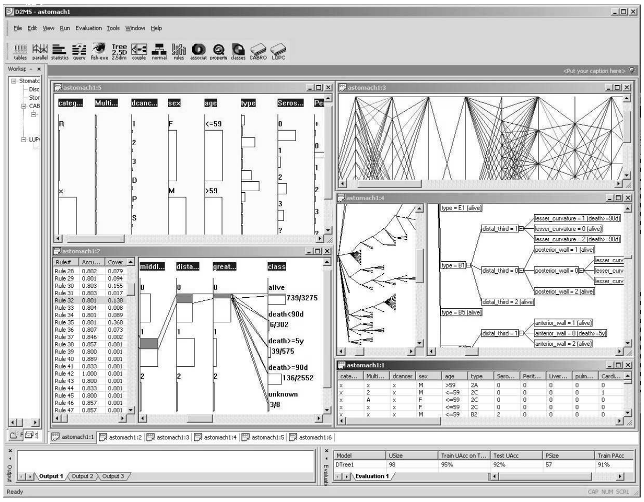
selection. The support consists of providing descriptive analysis of data, selection of appropriate algorithms their and parameters, and performance metrics to evaluate the findings [12].

Figure 9 shows a conceptual architecture of D2MS where the Data Mining module includes our decision tree learning method CABRO, and rule learning method LUPC. D2MS first supports the user in doing trials on combinations of algorithms and their parameter settings in order to produce competing models, and then it supports the user in evaluating them quantitatively and qualitatively by providing both performance metrics values as well as visualization of these models. In D2MS, visualization helps the user to interpret data and models as well their connections in order to understand and evaluate models better. Through visualization the user also can see better the effect of algorithm settings on resulted models so that he/she can adjust settings to reach adequate models.

There are three visualizers in D2MS: data visualizer, rule visualizer, and hierarchy visualizer [12], [24]. The data visualizer provides the user graphical views on the statistics of the input data and relations between attributes. These include multiple forms of viewing data such as tables, cross-tabulations, pie or charts, cubes. The data visualizer supports the user in the preprocessing step and in the selection of algorithms when registering plans (Figure 10).

The rule visualizer allows the user to view rules generated by CABRO or LUPC. The visualizer provides graphical views on statistics of conditions and conclusions of a rule, correctly and wrongly matched cases in both training and testing data, and links between rules and data. The hierarchy visualizer visualizes hierarchical structures generated by D2MS. The visualizer provides different views that may be suitable for different types and sizes of hierarchies. The user can view an overall structure of a hierarchy together with the detailed information of each node.

D2MS provides several visualization techniques that allow the user to visualize large hierarchical structures effectively. The tightly-coupled view simultaneously displays a hierarchy in normal size and tiny size that allows the user to determine quickly the field-of-view and to pan to the region of interest. The fish-eye



Rule#	Accu...	Cover
Rule 28	0.802	0.079
Rule 29	0.801	0.094
Rule 30	0.803	0.155
Rule 31	0.803	0.017
Rule 32	0.801	0.138
Rule 33	0.804	0.008
Rule 34	0.801	0.089
Rule 35	0.801	0.368
Rule 36	0.807	0.073
Rule 37	0.846	0.002
Rule 38	0.857	0.001
Rule 39	0.800	0.001
Rule 40	0.889	0.001
Rule 41	0.833	0.001
Rule 42	1.000	0.001
Rule 43	0.800	0.001
Rule 44	0.833	0.001
Rule 45	0.800	0.001
Rule 46	0.857	0.001
Rule 47	0.857	0.001

cate...	Multi...	dcancer	sex	age	type	Sero...	Perit...	Liver...	pulm...	Cardi...
x	x	x	M	>59	2A	0	0	0	0	0
x	2	x	M	<=59	2C	0	0	0	0	1
x	A	x	F	<=59	2C	0	0	0	0	0
x	x	x	F	<=59	2C	0	0	0	0	0
x	x	x	M	<=59	B2	2	0	0	0	0

Model	USize	Train UAcc on T...	Test UAcc	PSize	Train PAcc
DTree1	98	95%	92%	57	91%

Fig. 3 Data and knowledge visualization in D2MS

view distorts the magnified image so that the center of interest is displayed at high magnification, while the rest of the image is progressively compressed. Also, the new technique T2.5D (Trees with 2.5 Dimensions) is proposed and implemented in D2MS for visualizing very large hierarchical structures.

3. The Integrated Approach in Hepatitis Study

In this session we will show that the integration of three factors mentioned above in medical data mining allows us to obtain better results than working with them separately.

Viral hepatitis is a disease of the inflammation of the liver by the infection of hepatitis virus. As viral hepatitis has a potential risk to become liver cirrhosis and hepatocellular carcinoma (HCC) – which is the most common type of liver cancer – studies on viral hepatitis, specially on hepatitis type B and type C, are important in hepatitis study.

The hepatitis temporal database collecting from 1982 to 2001 at the Chiba university hospital was given to challenge the data mining community [21]. This database contains results of 983 laboratory tests of 771 patients. It is a large un-cleansed temporal relational database consisting of six tables of which the biggest has 1.6 million records:

T1. Basic information of patients (771 records)
T2. Results of biopsy (960 records)
T3. Information on interferon therapy (198 records)
T4. Measurements in in-hospital tests (459 records)
T5. Results of out-hospital medical tests (30,243 records)
T6. Results of in-hospital medical tests (1,565,877 records).

The doctors posed a number of problems on hepatitis that are expected to be investigated by KDD techniques. For last five years together with physicians we have been using our integrated approach to investigate this database in particular three significant problems in hepatitis study [21]:

P1. Discover the differences in temporal patterns between hepatitis B and C (HBV and HCV).
P2. Evaluate whether laboratory tests can be used to estimate the stage of liver fibrosis (LC (liver cirrhosis) vs. nonLC (non liver cirrhosis)).
P3. Evaluate whether the interferon therapy is effective or not.

Our data mining solution to exploit the hepatitis database is temporal abstraction (TA) that is an approach to temporal pattern detection aiming at deriving an abstract description of temporal data by extracting their most relevant features over periods of time [2], [4]. We approach the hepatitis database by novel temporal abstraction methods aiming at explaining the causes and mechanisms of hepatitis diseases in a comprehensible way to physicians.

We have developed two temporal abstraction methods, one is *abstraction pattern extraction* (APE) whose task is to map (to abstract) a given fixed length sequence into one of predefined abstraction patterns [13], [19], and the other is *temporal relation extraction* (TRE) whose task is to find temporal relations in terms of temporal logic among detected temporal patterns, and use these relations together with abstraction patterns to solve problems [11]. Our framework consists of four steps as shown in (Figure 11).

3.1 Temporal Relation Extraction

Typical TA works in the literature deal with regular temporal data, says, temporal data of an individual measured on consecutive days in a short period, diabetes data measured on consecutive days within two weeks; newborn infants regularly measured every minute [2]. However, the hepatitis data is characterized by irregular time stamped points in long periods [13].

Temporal logic [1] was developed as a theory of action and time by Allen whose basis is relations between temporal events (Figure 12). Recently, there are works on finding association rules based on temporal relations (see [2]).

1. Transforming the original hepatitis temporal database into a database of temporal patterns abstracted by the proposed temporal abstraction algorithms.
2. Using D2MS and other learning methods as CBA, C5.0, etc. to find rules on relations between temporal patterns from the abstracted database.
3. Exploiting MEDLINE for background or domain knowledge to support the knowledge evaluation.
4. Analyzing the findings with/by physicians.

Fig. 4 The framework for mining hepatitis data.

In this paper we illustrate the integrated approach by TRE data mining method. The key idea that makes TRE work differs from other methods of temporal pattern detection is domain-oriented temporal patterns are found by properties of hepatitis disease but not in a formal manner. The key steps of the TRE method include: (a) to detect temporal patterns from irregular temporal sequences and to softly find temporal relations among detected patterns by proposed algorithms, and (b) to use available data mining algorithms to find patterns/models [11].

3.2 Extracting Knowledge from MEDLINE

MEDLINE has a special potential for any of medical applications, for example, medical data mining. There are about 60,000 abstracts of research papers concerning hepatitis. Our target is to extract information/domain knowledge from MEDLINE abstracts on hepatitis study and use them in mining the hepatitis database. To this end, we have attempted to exploit MEDLINE in two ways.

The first is, for the filtering purpose in preprocessing, to find strong associations of the laboratory tests that can be used as selected features to narrow the search space in the mining algorithms. However, a crucial problem occurred when searching associations of laboratory tests from the set of MEDLINE abstracts concerning hepatitis, which are usually short, is that often no such associations of the tests exist in MEDLINE. This problem is not only happened in hepatitis study but can also in other fields. Our solution is to use tolerance rough set model (TRSM) [10], [17] that represents the bag of words (BOW) of each abstract by its TRSM upper approximation, i.e., each word in the abstract BOW is surrogated by all words in its tolerance class. Thus, the TRSM upper representation of the abstract contains more words than the abstract BOW and this representation gives more chances to find strong combinations of significant laboratory tests, and the problem caused by the small size of abstracts is expected to be solved.

The advantages of dealing with text by TRSM-based association mining are: (i) it gives a meaningful interpretation in the context of information retrieval about the dependency and the semantic relation of index terms; and (ii) it is relatively simple and computationally efficient. Association rule mining is also known as an efficient way to find strong combinations among items. To detect the strong relationships among blood tests appearing in MEDLINE abstracts we employ those two methods in the following manner [17]

- To approximately find a surrogate for each MEDLINE abstract by its TRSM upper representation

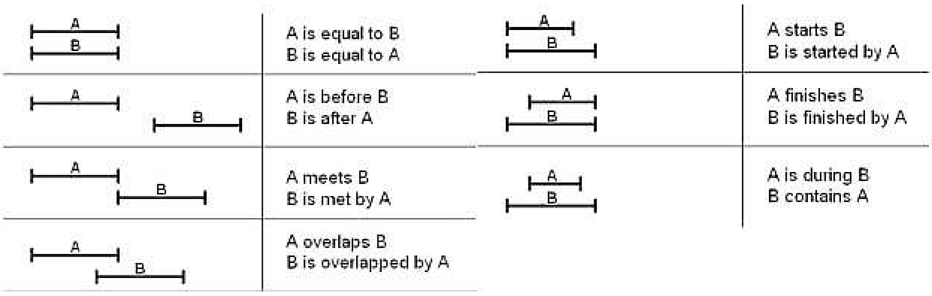


Fig. 5 Temporal relations in Allen's temporal logic

that extends the part of interest in the MEDLINE abstract.

- To find associations of laboratory tests from the set of surrogates of MEDLINE abstracts.

The second is to interpret and evaluate rules found in the post processing by exploiting discovered knowledge in MEDLINE abstracts. For each rule found we have tried to find related discovery from other works reported in MEDLINE that either conflict or agree with our finding. There is no way to find exactly article abstracts that related to discovered rules but we can filter for a small set of abstracts that potentially related then manually check them. This is done by using the temporal patterns appeared in the rule as query for TRSM-based information retrieval [10]. Note that though academic papers tend to have similar formats and similar terminology, but still there is a various usage of terms. We have unified those synonyms based on *The Medical Subject Headings*(MeSH) thesaurus for MEDLINE, then specially indexed the abstracts for our purpose.

3.3 Knowledge Visualization

We illustrate the use of rule visualization in support interpreting and understanding discovered rules in hepatitis study by our APE method [14], in particular the LC vs. non-LC problem.

Figure 13 shows a typical screenshot of rule visualizer in doing this task. When using default parameters of program LUPC in D2MS (90% minimum confidence and 5 as minimum support count), we found 1 rule for LC and 29 rules for non-LC. Assume that we want to assess the rule's interestingness, for example, rule #4 of non-LC whose precedent is the conjunction of "TTT_High = Y", "GOT_High = Y", "LDH_NormalToLow = Y" and "T-BIL_NormalToLow = Y" as shown in Figure 13.

Rules of each class are displayed with a color, concretly red for LC class and dark green for non-LC class, while attribute-value pairs are displayed with yellow color. The rule "#4 non-LC" is shown at the center in connecting with its four attribute-value pairs in the precedent part. Each attribute-value pair then is linked to all rules where it appears in the precedent parts. In this case, we can easily observe that temporal patterns "TTT_High", "GOT_High", "DH_NormalToLow" and "T-BIL_NormalToLow" occurred only in non-LC rules. By a double click on the attribute-value node "TTT_High" we can see in another screen the links from this node to all the rules that contains "TTT_High" in their precedent parts, and other conditions in the precedent part of each rule (Figure 14). In fact, we can switch between these two modes in rule visualizer to observe each rule or attribute-value pair and its related conditions or rules, i.e., its neighborhood information.

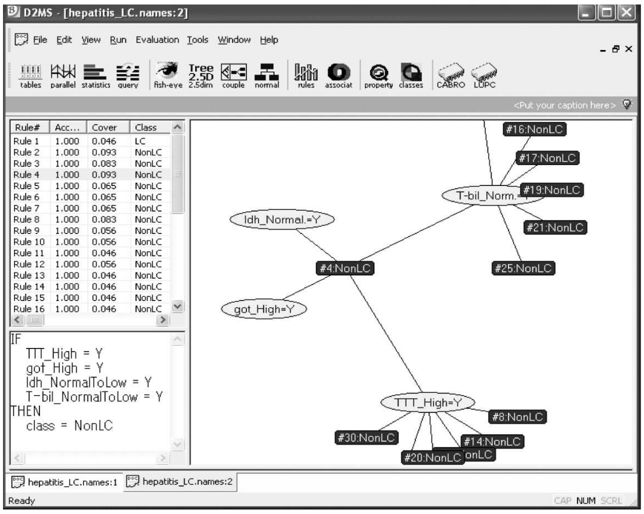
Various rules found by D2MS with its visualization tools from stomach cancer and hepatitis data have been encouragingly evaluated by physicians and they are really interested in the tools, especially, in case to compare the impacts of similar content of rules.

3.4 Findings and Evaluation

For the problem P1 of finding differences in temporal patterns between HBV and HCV, we created an abstracted dataset from original data of 584 patients (225 of HBV and 359 of HCV). Program CBA found totally 317 rules of which hypothesis testing accepts 226 rules with statistical significance (118 rules for HBV and 108 for HCV). Program LUPC found totally 300 rules, and 113 remained after hypothesis testing (71 for HBV and 42 for HCV). C5.0 found 29 rules for HBV (default is HCV). Some results evaluated in connection with MEDLINE abstracts and visualization were presented below.

Findings which are different from clinical observation.

The physicians suggested that in their clinical observation there is no clear distinction between HBV and HCV. However, when analyzing the rules we observed the followings: There are various rules each matches a considerable number of patients for either HBV or HCV with rather high confidence. Figure 14 presents some obtained rules, for example:

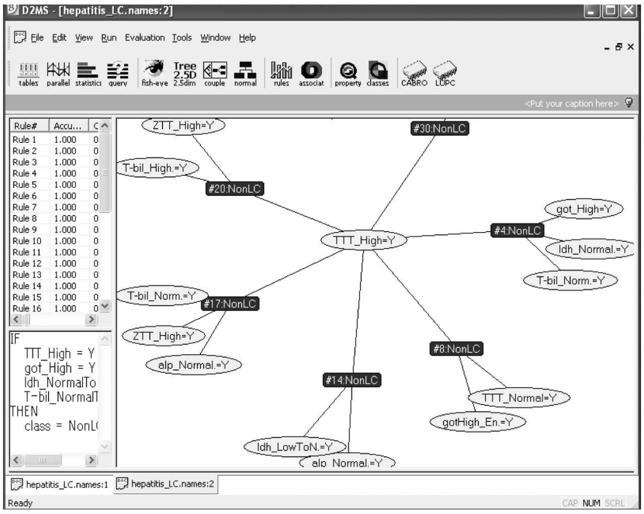


D2MS - [hepatitis_LC.names:2]

Rule#	Acc...	Cover	Class
Rule 1	1.000	0.046	LC
Rule 2	1.000	0.093	NonLC
Rule 3	1.000	0.083	NonLC
Rule 4	1.000	0.093	NonLC
Rule 5	1.000	0.065	NonLC
Rule 6	1.000	0.065	NonLC
Rule 7	1.000	0.065	NonLC
Rule 8	1.000	0.083	NonLC
Rule 9	1.000	0.056	NonLC
Rule 10	1.000	0.056	NonLC
Rule 11	1.000	0.046	NonLC
Rule 12	1.000	0.056	NonLC
Rule 13	1.000	0.046	NonLC
Rule 14	1.000	0.046	NonLC
Rule 15	1.000	0.046	NonLC
Rule 16	1.000	0.046	NonLC

```
IF
   TTT_High = Y
   got_High = Y
   ldh_NormalToLow = Y
   T-bil_NormalToLow = Y
THEN
   class = NonLC
```

Fig. 6 Visualization of rule 4 *NonLC* in detail



D2MS - [hepatitis_LC.names:2]

Rule#	Accu...	C
Rule 1	1.000	0
Rule 2	1.000	0
Rule 3	1.000	0
Rule 4	1.000	0
Rule 5	1.000	0
Rule 6	1.000	0
Rule 7	1.000	0
Rule 8	1.000	0
Rule 9	1.000	0
Rule 10	1.000	0
Rule 11	1.000	0
Rule 12	1.000	0
Rule 13	1.000	0
Rule 14	1.000	0
Rule 15	1.000	0
Rule 16	1.000	0

```
IF
   TTT_High = Y
   got_High = Y
   ldh_NormalTo
   T-bil_NormalT
THEN
   class = NonL(
```

Fig. 7 A rule viewed in its relations with others.

R#2 [HCV]: “TTT in high state with peaks” AFTER ”ZTT in high state with peaks” (support count = 86, conf. = 0.73).

R#8 [HBV]: “GOT in very high state with peaks” ENDS ”GPT in extreme high state with peaks” (support count = 41, conf. = 0.71).

These rules show different temporal relation-based patterns that distinguish well groups of HBV and HCV patients. Some typical differences in HBV and HCV can be seen in temporal patterns concerning state changes of ALP (alkaline phosphatase) follows that of LDH (lactate dehydrogenase). Program CBA extracts 7 rules with this type of relation with high support and confidence, for example:

R#13 [HBV]: “ALP changed from low to normal state” AFTER ”LDH changed from low to normal state” (support count = 21, conf. = 0.71).

R#5 [HCV]: “ALP changed from normal to high state” AFTER ”LDH changed from normal to low state” (support count = 60, conf. = 0.80).

In relation to state changes in ALP, we found rules covering the cases where ALP moves between high and normal states in HCV, and between normal and low states in most of HBV. The only exceptional rule in HBV that covers 3 cases is

R#201 [HBV]: “ALP changed from normal to high state” ENDS ”LDH changes from low to normal state” (support count = 38, conf. = 0.80).

Findings in comparison with Medline abstracts.
When matching our temporal relation-based results against medical knowledge, we look for reported results on medical journals from Medline abstracts. There are two papers that our results reconfirm their findings within the temporal relation setting. The work of [23] showed that the main difference between HBV and HCV is that the base state of TTT (thymol turbidity test) in HBV is normal, while that of HCV is high. In a temporal relation with ALB (albumin) changing from normal to low, TTT is normal for HBV but high for HCV. This shows that relations respect the reported knowledge that TTT is higher for HCV.

R#172 [HBV]: “ALB changed from normal to low state” BEFORE ”TTT in normal state with peaks” AND ”GPT in high state with peaks” AFTER ”TTT in normal state with peaks” (support count = 5, conf. = 1.00).

R#53 [HCV]: ”ALB changed from normal to low state” BEFORE ”TTT in high state with peaks” AND “ALP from normal to high state” BEFORE ”TTT in high state with peaks” (support count = 10, conf. = 1.00).

R#83 [HCV]: “ALB changed from normal to low state” AFTER “TTT in high with peaks” AND “TTT in high state with peaks” AFTER “ZTT in very high state with peaks” (support count = 8, conf. = 1.00).

The work of [25] stated that for HBV cases, LAP (leucine aminopeptidase) is modified without any irreversible damage to the corresponding organs. The rules reported below show that for HBV, there is no evidence of other peaks in TTT or so after LAP changes. On the other hands, in HCV cases, there are many evidences that inflammation occurs (TTT shows peaks) after the change in LAP state. The temporal relation-based patterns also prove that in HBV cases, there is no irreversible damage (no inflammations) as LAP changes its state, but not for HCV cases. For example:

R#137 HBV: “LAP changed from high to normal state” BEFORE “LDH changed from low to normal state” AND “GPT in high state with peaks" AFTER "TTT in normal state with peaks" (support count = 5 conf. = 1.00).

R#62 HCV: "LAP changed from high to normal state" AFTER "TTT in high state with peaks" AND "GPT in very high state with peaks" BEFORE "TTT in high state with peaks" (support count = 9 conf. = 1.00).

From the analysis, we can conclude that the temporal relation rules conform to some reported medical knowledge. The temporal relation itself can show the different effects between HBV and HCV when a liver product changes its state.

Rules on hepatitis were found and evaluated not only with support of MEDLINE abstract mining but also with active participation of physicians. Some of them are especially specialized in hepatitis and some play roles of intermediators between their hepatitis specialist and computer scientists.

As shown in this paper, data mining algorithms produce a number of patterns/rules to which there is room for different interpretations when they are provided in an isolated list. Only hepatitis specialists can give deeper views and new directions about them with support of translation by our intermediators. In addition they also suggest the possibilities that some other features/information we have not included in mining may have strong relations to obtained rules, such as sex, age range and patient history. Our visualization tools, as shown in the previous section, succeed to provide them to capture the importance of some specific features more than the others and the relations among feature groups and patients. Finally some of recent rules listed below were judged to be potentially new and useful.

R#10 NonLC: "GPT in very high state with peaks" AFTER "TTT in high state with peaks" AND "GOT in very high state with peaks" ENDS "GPT in very high with peaks" AND "GOT in very high state with peaks" AFTER "TTT in high state with peaks" (support count = 10, conf. = .80).

R#8 LC: "GPT in very high state with peaks" AFTER "TTT in very high state with peaks" AND "GPT in very high state with peaks" BEFORE "TTT in high state with peaks" AND "GOT in very high state with peaks" AFTER "TTT in high state with peaks", (support count = 8, conf. = .80).

R#42 NonLC: "LDH changed from normal to low state" BEFORE "TTT in normal state with peaks" AND "GOT in very high state with peaks" BEFORE "ZTT in high state with peaks" AND "GOT in very high state with peaks" AFTER "GPT in very high state with peaks" (support count = 6, conf. = .86).

For example, the last one interests the physicians because of its unexpectedness coming from their common knowledge that CRE (creatinine) indicates kidney conditions and does not directly suggest hepatic disorders, to which recently some papers have shown the relations to hepatitis [26], although it still requires further detailed analysis and investigations.

4. Conclusion

We have proposed an integrated approach that merges data mining and text mining methods plus visualization support for expert evaluation and applied it in the hepatitis study. Recent results in our practice are of potentially new and interesting. Besides the physicians' verifications of those obtained knowledge in order to really discover the new and interesting medical knowledge, we need to consider to brush up our integrated approach, especially providing more various visualizations of additional information from many directions and develop other types of domain knowledge extraction from medical literature.

This integrated approach in medical data mining can also be applied to other fields in medicine.

5. Acknowledgements

This work is partially supported by Grant-in-Aid for Basic research (B), the Ministry of Education, Science, Sports and Culture. The authors would like to express their sincere thanks to Drs. Hideto Yokoi, Nguyen Trong Dung, Nguyen Duc Dung, Le Si Quang for their collaboration.

References

[1] Allen, J., "Maintaining Knowledge About Temporal Intervals", *Communications of the ACM, 26(11)*, 832–843, 1983.

[2] Balaban, M., Boaz, D., and Shahar, Y., "Applying temporal abstraction in medical information systems", *Annals of mathematics, computing and teleinformatics 1(1)*, 56-64, 2003.

[3] Chittaro, L., "Information Visualization and Its Application to Medicine", *Artificial Intelligence in Medicine 22*, 81–88, 2001.

[4] Chittaro, L., Montanari, A., "Temporal representation and reasoning in artificial intelligence: Issues and Approaches", *Annals of Mathematics and Artificial Intelligence 28*, 47–106, 2000.

[5] Cios, K.J., *Medical data mining and knowledge discover* (Ed.), Physica-Verlag, 2001.

[6] Cios, K.J., Moore, G.W., "Uniqueness of medical data mining", *Artificial Intelligence in Medicine, 26*, 1–24, 2002.

[7] Cohen, A.M., Hersh, W.R. "A Survey of Current Work in Biomedical Text Mining", *Briefings in Bioinformatics*, , 6(1) 57–71, 2005.

[8] Fayyad, U.M., Grinstein. G.G., Wierse, A. *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kauffmann, 2002.

[9] Hirechman, L., Park, J.C.,Tsujii, J., Wong, L., Wu, C.H., "Accomplishments and challenges in literature data mining for biology", *Bioinformatics*, 18(12), 1552-1561, 2002.

[10] Ho, T.B., Kawasaki, S., Nguyen, N.B., "Cluster-based Information Retrieval with a Tolerance Rough Set Model", *International Journal of Fuzzy Logic and Intelligent Systems*, KFIS, Vol. 2, No. 1, 26–32, 2002.

[11] Ho, T.B., Nguyen, C.H., Kawasaki, S., Takabayashi, K., "Exploiting Temporal Relations in Mining Hepatitis Data", *Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD'07*, LNAI, Springer, 2007 (in press).

[12] Ho, T.B., Nguyen, T.D., Shimodaira, H., Kimura, M., "A Knowledge Discovery System with Support for Model Selection and Visualization", *Applied Intelligence*, Kluwer Academic Publishers, Vol. 19, Issue 1-2, 125–141, 2003.

[13] Ho, T.B., Nguyen, T.D., Kawasaki, S., Le, S.Q., Nguyen, D.D., Yokoi, H., Takabayashi, K. , "Mining Hepatitis Data with Temporal Abstraction", *ACM International Conference on Knowledge Discovery and Data Mining KDD'03*, 369–377, 2003.

[14] Ho, T.B., Nguyen, T.D., Kawasaki, S., "Failures and Successes in Medical Data Mining", *Chapter 6, Knowledge-Based Intelligent Systems for Health Care*, T. Ichimura and Yoshida, K. (Eds.), Advanced Knowledge International Publishers, 167–212, 2004.

[15] Ho, T.B., Saito, A., Kawasaki, S., Nguyen, D.D., Nguyen, T.D., "Failure and Success Experience in Mining Stomach Cancer Data", , *International Workshop Data Mining Lessons Learned*, ICML'02, 41–47, 2002.

[16] Kononenko, I., "Learning dependencies in multivariate times series", *tArtificial Intelligence in Medicine, 23*, 89–109, 2001.

[17] Kawasaki, S., Ho, T.B.,"Extracting background knowledge from medical literature", *Symposium on Data/Text Mining from Large Databases, IFSR'05*, S5-4-3, 2005.

[18] Kawasaki, S., Ho, T.B., "An Integrated Approach in Medical Data Mining", *The First International Conference on Knowledge, Information and Creativity Support Systems KICSS'06*, Ayutthaya, 24–31, 2006.

[19] Kawasaki, S., Nguyen, T.D., Ho, T.B.,"Temporal Abstraction for Long-Term Changed Tests in the Hepatitis Domain", *Journal of Advanced Computational Intelligence & Intelligent Informatics*, 17(3):348–354, 2003.

[20] Kawasaki, S., Saitou, A., Nguyen, D.D., Ho, T.B., "Mining from Medical Data: Case-Studies in Meningitis and Stomach Cancer Domains", *KES-02 6th International Conference on Knowledge-based Intelligent Information & Engineering Systems*, 547–551, 2002.

[21] http://lisp.vse.cz/challenge/ecmlpkdd2004/

[22] Lavrak, N., "Selected techniques for data mining in medicine", *Artificial Intelligence in Medicine*, 16, 2–23, 125–134, 1999.

[23] Murawaki Y., Ikuta Y., Koda M., Kawasaki H., "Comparison of clinical laboratory liver tests between asymptomatic HBV and HCV carriers with persistently normal aminotransferase serum levels", *Hepatol Research 21(1)*, 67–75, 2001.

[24] Nguyen, D.D., Ho, T.B., Kawasaki, S., "Knowledge Visualization in Hepatitis Study", *Asia Pacific Symposium on Information Visualization APVIS'06*, 59–62, 2006.

[25] Sakai H., Horinouchi H., Masada Y., Takeoka S., Ikeda E., Takaori M., Kobayashi K., Tsuchida E., "Metabolism of hemoglobin-vesicles (artificial oxygen carriers) and their influence on organ functions in a rat model" *Biomaterials 25(18)*, 4317–4325, 2004.

[26] Sezer, S., "Hepatitis C infection in hemodialysis patients: Protective against oxidative stress?", *Transplant Proc.* 38(2):406–410, 2006.

Tu Bao Ho is a professor at School of Knowledge Science, Japan Advanced Institute of Science and Technology, Japan. He received his M.S. and Ph.D. from Marie and Pierre Curie University in 1984 and 1987, respectively. His research interest include knowledge-based systems, machine learning, data mining, medical informatics and and bioinformatics.

Saori Kawasaki is a research associate at School of Knowledge Science, Japan Advanced Institute of Science and Technology (JAIST), Japan. She received her M.S. and Ph.D. from JAIST in 2000 and 2003, respectively. Her research interest includes data mining, knowledge evaluation and medical informatics.

Katsuhiko Takabayashi is MD, FACP and professor at Division for Medical Informatics and M anagement, Chiba University Hospital, 1-8-1 Inohana, Chuoku, Chiba, 260-8677, Japan.

Canh Hao Nguyen is a PhD student at School of Knowledge Science, Japan Advanced Institute of Science and Technology, Japan. He received his Bachelor of Science in Computer Science (with Honours) from the University of New South Wales in 2002. He research interests lie in statistical machine learning, ranging from theoretical to application studies.

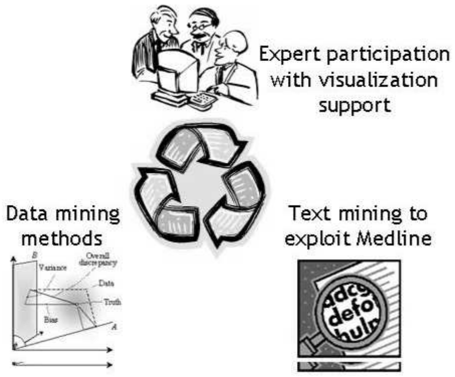


Fig. 8 Three factors of the integrated approach.

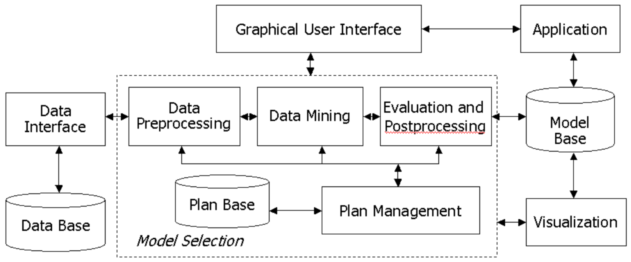


Fig. 9 Conceptual architecture of D2MS

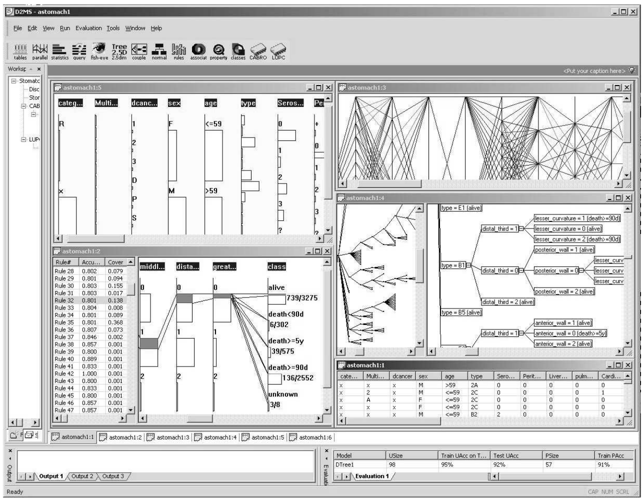
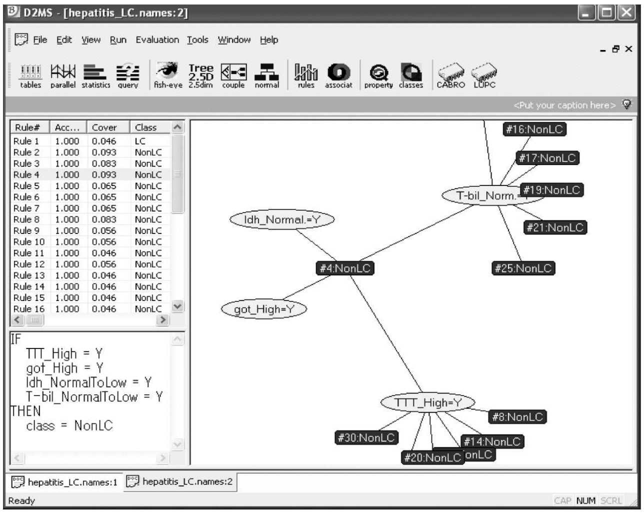


Fig. 10 Data and knowledge visualization in D2MS

1. Transforming the original hepatitis temporal database into a database of temporal patterns abstracted by the proposed temporal abstraction algorithms.
2. Using D2MS and other learning methods as CBA, C5.0, etc. to find rules on relations between temporal patterns from the abstracted database.
3. Exploiting MEDLINE for background or domain knowledge to support the knowledge evaluation.
4. Analyzing the findings with/by physicians.

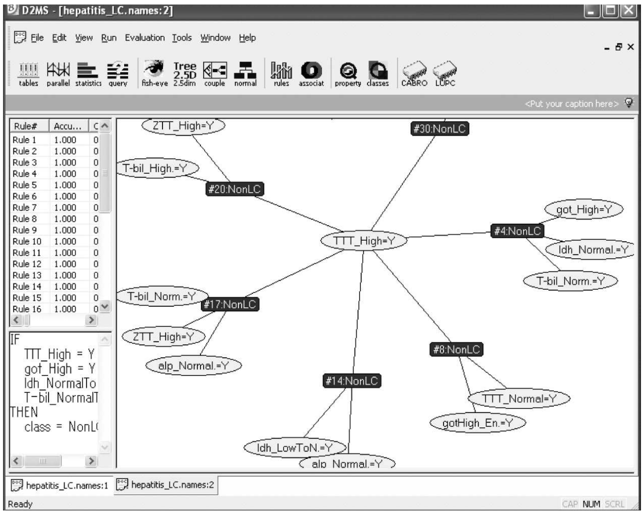
Fig. 11 The framework for mining hepatitis data.



Rule#	Acc...	Cover	Class
Rule 1	1.000	0.046	LC
Rule 2	1.000	0.093	NonLC
Rule 3	1.000	0.083	NonLC
Rule 4	1.000	0.093	NonLC
Rule 5	1.000	0.065	NonLC
Rule 6	1.000	0.065	NonLC
Rule 7	1.000	0.065	NonLC
Rule 8	1.000	0.083	NonLC
Rule 9	1.000	0.056	NonLC
Rule 10	1.000	0.056	NonLC
Rule 11	1.000	0.046	NonLC
Rule 12	1.000	0.056	NonLC
Rule 13	1.000	0.046	NonLC
Rule 14	1.000	0.046	NonLC
Rule 15	1.000	0.046	NonLC
Rule 16	1.000	0.046	NonLC

```
IF
   TTT_High = Y
   got_High = Y
   ldh_NormalToLow = Y
   T-bil_NormalToLow = Y
THEN
   class = NonLC
```

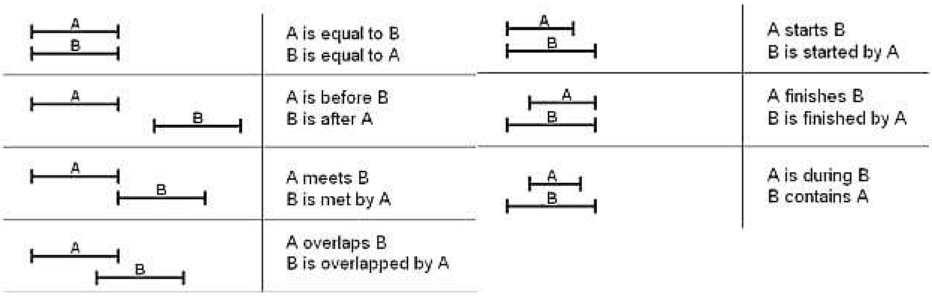
Fig. 13 Visualization of rule $4NonLC$ in detail



Rule#	Accu...	C
Rule 1	1.000	0
Rule 2	1.000	0
Rule 3	1.000	0
Rule 4	1.000	0
Rule 5	1.000	0
Rule 6	1.000	0
Rule 7	1.000	0
Rule 8	1.000	0
Rule 9	1.000	0
Rule 10	1.000	0
Rule 11	1.000	0
Rule 12	1.000	0
Rule 13	1.000	0
Rule 14	1.000	0
Rule 15	1.000	0
Rule 16	1.000	0

```
IF
   TTT_High = Y
   got_High = Y
   ldh_NormalTo
   T-bil_NormalT
THEN
   class = NonL(
```

Fig. 14 A rule viewed in its relations with others.



A is equal to B
B is equal to A

A is before B
B is after A

A meets B
B is met by A

A overlaps B
B is overlapped by A

A starts B
B is started by A

A finishes B
B is finished by A

A is during B
B contains A

Fig. 12 Temporal relations in Allen's temporal logic