JAIST Repository
https://dspace.jaist.ac.jp/

Title	Extracting Background Knowledge from the Medical Literature
Author(s)	Kawasaki, Saori; Tu, Bao Ho
Citation	
Issue Date	2005-11
Type	Conference Paper
Text version	publisher
URL	http://hdl.handle.net/10119/3916
Rights	©2005 JAIST Press
Description	The original publication is available at JAIST Press http://www.jaist.ac.jp/library/jaist-press/index.html, IFSR 2005 : Proceedings of the First World Congress of the International Federation for Systems Research : The New Roles of Systems Sciences For a Knowledge-based Society : Nov. 14-17, 2126, Kobe, Japan, Symposium 5, Session 4 : Data/Text Mining from Large Databases Text Mining

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Extracting Background Knowledge from the Medical Literature

Saori Kawasaki and TuBao Ho

School of Knowledge Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
skawasa@jaist.ac.jp, bao@jaist.ac.jp

ABSTRACT

The growing interest in dealing with the information in medical publications is not only because of the scene of physicians' medical practices as the evidence-based medicine (EBM) but also encouraged by the realization of computational approaches to the electrically stored medical data, such as data mining and many attempts to enrich them. Among many different purposes of means of utilizing MEDLINE, our target is to extract background knowledge from MEDLINE abstracts, and exploit it in the mining from a certain medical database. Generally, information extraction approaches try to understand the contents in text, and extract the appropriate information to the requests. While those requires complex and time consuming procedures to complete, we propose a method for extracting background knowledge useful for data mining from medical database in a simple but effective manner. Main idea of this method is to find strong combinations among clinical test items in order to use the result for feature selection and narrowing the mined result. We design the framework based on rough set theory for considering the word ambiguity. This framework consists of three steps: to approximately represent MEDLINE abstracts related to the topic under investigation by the tolerance rough set model, then to detect associations between terms related to the topic, and finally to discover rules with our rule learning program LUPC when considering found associations as input to exclusive or inclusive constraints of LUPC. Some sets of parameters succeed to extract plausibly useful combinations of clinical test items for finding patterns from the hepatitis database.

Keywords: text mining, rough sets, MEDLINE.

1. INTRODUCTION

Along with rapid increase of avairable scientific literature in electric forms, methods to process such a rich scientific source, such as information retrieval, text summarization, information extraction or text mining specially for the scientific text collections, have been proposed in order to suppoort users who are interested in those authorized specific sources. MEDLINE is one of most famous literature databases of the biomedical and life science [1], and also supports the trend and practices of the evidence-based medicine (EBM).

Therefore, MEDLINE has a special potencial for any of the medical applications, for example, data mining in medicine. Among many different purposes to process MEDLINE [2, 3], our target is to extract information/background knowledge from MEDLINE abstracts and exploit the background information in mining medical databases as in Fig. 1.

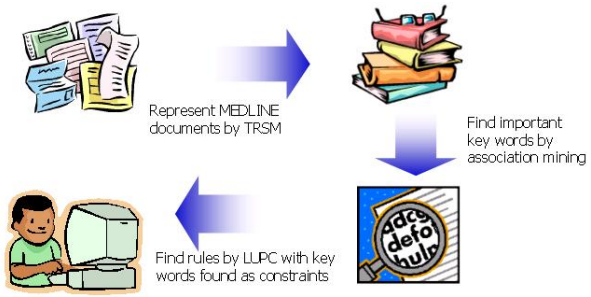

Fig. 1: The basic idea of rule induction with background knowledge extracted from medical literature

Our motivation is the fact that many of data mining techniques produce too many patterns from the target database. Generally it becomes a tough work for users to find anything interesting among them, as most of them are trivial or of low reliability, and also as the number of pattens is too many for users to directly evaluate. Therefore, it is useful to extract information from MEDLINE and use them to narrow the feature space of original data to search, or to filter the mined patterns in order to provide users/evaluators less number of more plausible patterns.

The objectives of this work is to propose a method to extract sets of combinations of medical examinations in terms of Hepatitis. Information such as combinations of

attibutes, as learning constraints of LUPC [4], can be used in the mining hepatitis database,provided by Chiba University Hospital, Japan [5], and be expected to contribute to induct more plausive rules than general trials. We developed a technique to find surrogates of MEDLINE abstracts with approximations of topics of interest by tolerance classes, then detect topic associations from the surrogates base on Apriori, the most famous assocaition algorithm [6].

In this paper, we briefly introduce background including our target database, temporal abstraction as data tranformation for it, LUPC as data mining system to which background knowledge from MEDLINE aims to contributes, and rough sets models for text as the key technique to represent abstracts in MEDLINE in section 2. Section 3 describes the framework for extracting associations among medical examinations based on tolerance rough set model and association mining. Section 4 reports the experimental results of this approach, then we conclude our work in section5.

2. BACKGROUND

2. 1. Introduction to the Target Database

Hepatitis database was provided as a common database with goals posed by the data provider, Chiba University Hospital, Japan, at the active mining project [7] and at the discovery challenge of PKDD 2002-2004 [5].

Hepatitis Database
This database had been collected by Chiba University Hospital during 1981 and 2001 including 771 patients and 983 internal/external hospital examinations. These information are structured as a relational database consisting of six main tables of (a) Basic information of patients (b) Results of biopsy, (c) Information on interferon therapy, (d) Results of out-hospital, (e) Results of in-hospital examinations and (f) Results on hematological analysis. For table (g) it is proveided the normal range values (Information about measurements in in-hospital examinations). By joining tables with patient ID as key, each patient can be described in a set of time series of examinations.

Posed goals to the hepatitis database
Originally, six or four gaols are posed for the database as in [5, 7]. Our first step focuses on these three problems among them.

(1) Discover the differences in temporal patterns between hepatitis B and C.
(2) Evaluate whether laboratory examinations can be used to estimate the stage of liver fibrosis. (If possible, we may be able to use laboratory examinations as the substitutes for biopsy because biopsy is invasive to patients.)
(3) Evaluate whether the interferon therapy is effective or not.

2. 2. Temporal Abstraction

The original database has a complex relational database structure and time related information. All of posed problems to the database expect to have classification patterns of patient groups. For those targets, it is convenient to design data representations to mine focusing on patient.

Temporal abstraction (TA) has been attempted to the medical time series in order to capture the essential and abstracted descriptions from time series data appropriate for medical usage, and the target datasets of typical researches done by [8] [9] are that of regularly collected or collected in short period. Different from them, TA approach by [10] considers the characteristics of its target database, in which a patient's history of medical examination results is stored as multiple records of irregularly time stamped data for each examination item during long period.

The key idea is to transform a series of time-stamped data values by abstraction into an interval-based representation of data. For that purpose, [10] proposed TA framework consisting of two phases: basic TA that concerns with abstracting time-stamped data within episodes, and complex TA that concerns with temporal relationships between findings from a basic TA or from other complex TA (see Fig. 2). The complex TA step is almost equivalent to the data mining step on categorical dataset, in which the background knowledge extracted from MEDLINE can contribute.

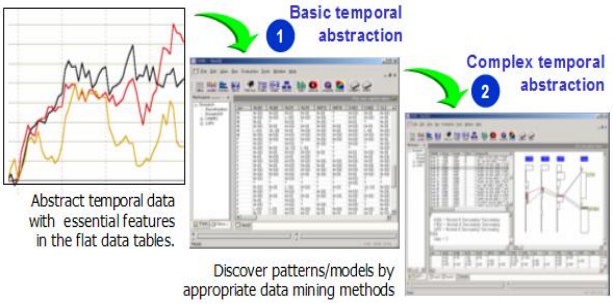

Fig. 2: Overview of Temporal Abstraction for the Hepatitis database

2. 3. Mining Hepatitis Data by using LUPC

LUPC was developed for leaning minority classes as accurate as majority classes even if the class distribution is very imbalanced [4]. It provides several effective parameters for learning, The most useful parameters which allow user to reflect their interst in the mining result are exclusive and inclusive constraints. Users can specify any attributes and/or values of a dataset as those constraints for learning at the left pane of LUPC as in Figyure 4, where values "?" and "O" of attribute "ALB", and attribute "ICG-15" are dropped off.

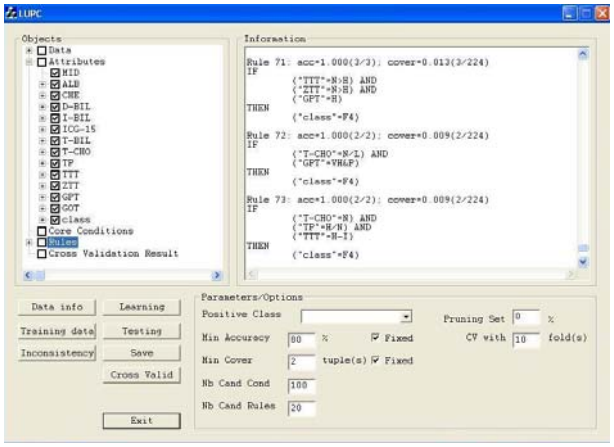

Figure 4. A screenshort of LUPC

These parameter settings may change the rules to be obtaind. For example, with conditions of 85% of minimum accuracy requirement and 10 cases for minimum coverage requirement, LUPC produces 44 rules without any constraints specifying, while it find 34 rules with same conditions but with specifying exclusive constraints which require to ommit missing values from all attributes.

2. 4. Tolerance Rough Sets Model for Text

2.4.1. Rough Sets

Rough set theory was introduced by Pawlak in early 1980s as a mathematical tool to deal with vagueness and uncertainty [11]. It has been applied in many fields such as machine learning, knowledge acquisition, decision analysis, information retrieval, pattern recognition, and recently knowledge discovery and data mining [12].

The starting point of theory of rough sets is that each set X in a universe U of objects can be "viewed" approximately by its *upper and lower approximations* in an equivalence space $\mathcal{R} = (U, R)$ where $R \subseteq U \times U$ is an equivalence relation, satisfying three properties of reflective (xRx), symmetric ($xRy \rightarrow yRx$) and transitive ($xRy \wedge yRz \rightarrow xRz$) for all $x, y, z \in U$. The lower and upper approximations in R of any subset X of U, denoted respectively by $\mathcal{L}(R, X)$ and $\mathcal{U}(R, X)$, are defined by

$$\mathcal{L}(R, X) = \{x \in U : [x]_R \wedge X \subseteq X\}$$

$$\mathcal{U}(R, X) = \{x \in U : [x]_R \wedge X \neq \varnothing\}$$

where $[x]_R$ denotes the equivalence class of objects indiscernible with x regarding the equivalence relation R. Then we call it *equivalence rough set model* (ERSM).

2.4.2. Tolerance Rough Sets Model for Text

As analyzed in [13], ERSM is not suitable for information retrieval and text processing due to the fact that the requirement of the transitive property in equivalence relations is too strict to the meaning of words, and there is no way to calculate automatically equivalence classes of terms, then [13] proposed a *tolerance rough set model* (TRSM) for text processing that adopts tolerance classes instead of equivalence classes as *tolerance relations* requires only *reflexive* and *symmetric* properties and allows overlapping classes.

Based on a general approximation model using *tolerance relations* introduced in [12], [13] proposed a TRSM for text based on "co-occurrence of terms", with these reasons: it (i) gives a meaningful interpretation in the context of information retrieval about the dependency and the semantic relation of index terms; and (ii) is relatively simple and computationally efficient. This model defines the uncertainty function I depending on a threshold θ of co-occurrence as

$$I_\theta(t_i) = \{t_j \mid f_{\mathcal{D}}(t_i, t_j) \geq \theta\} \cup \{t_i\}$$

where $f_{\mathcal{D}}(t_i, t_j)$ means the number of documents in $\mathcal{D}$ in which two index terms t_i, and t_j co-occur. Then, finally, we obtained the tolerance space $\mathcal{R} = (\mathcal{T}, I, \nu, P)$ in which the lower approximation $\mathcal{L}(R, X)$ and the upper approximation $U(R, X)$ in $\mathcal{R}$ of any subset $X \subseteq \mathcal{T}$ can be defined as

$$\mathcal{L}(R, X) = \{t_i \in \mathcal{T} \mid \nu(I_\theta(t_i), X) > 0\}$$

$$U(R, X) = \{t_i \in \mathcal{T} \mid \nu(I_\theta(t_i), X) = 1\}$$

where $\mathcal{T}$ is the set of all index terms, ν is the vague inclusion function and $P: I(U) \rightarrow \{0,1\}$ is a structurality function.

2.4.2. Text Representations with Weighting

Each document d_j is represented by a set of index term t_i (e.g., keywords), and each term t_i in d_j may have different significance in relations to d_j and $\mathcal{D}$. To reflect the importance of t_i in d_j, d_j is represented as a weighted vector of terms, i.e., $d_j = (t_1, w_{1j}; t_2, w_{2j}; \ldots; t_r, w_{rj})$. The set of all index terms from $\mathcal{D}$ is denoted by $\mathcal{T} = \{t_1, t_2, \ldots,$

t_N,}. The term-weighting method here defines weights for terms in the upper approximation $\mathcal{U}(\mathcal{R}, d_j)$ of d_j. It ensures that each term in the upper approximation of d_j but not in d_j has a weight smaller than the weight of any term in d_j.

$$w_{ij} = \begin{cases} (1+\log(f_{d_j}(t_i)))\times\log\dfrac{M}{f_{\mathcal{D}}(t_i)} & \text{if } t_i \in d_j \\ \min\limits_{t_h \in d_j} w_{hj} \times \dfrac{\log(M / f_{\mathcal{D}}(t_i))}{1+\log(M / f_{\mathcal{D}}(t_i))} & \text{if } t_i \in \mathcal{U}(\mathcal{R}, d_j) \setminus d_j \\ 0 & \text{if } t_i \notin \mathcal{U}(\mathcal{R}, d_j) \end{cases}$$

The vector length normalization is applied to the upper approximation $\mathcal{U}(\mathcal{R}, d_j)$ of d_j. Note that the normalization is done when considering a given set of terms. Figure 5 illustrates a MEDLINE abstract and its TRSM representation described by id and weight numbers.

MED_1: correlation between maternal and fetal plasma levels of glucose and free fatty acids correlation coefficients … level at delivery is only slightly dependent upon the maternal level.

MED_1: 21-0.178679, 44-0.094230, 48-0.228942, 57-0.235588, 110-0.257558, 198-0.328567, 299-0.126899, 403-0.371317, 437-0.136658, 683-0.306114, 692-0.306114, 694-0.306114, 1840-0.289422, 2546-0.189904, 4546-0.321535.

Figure 5: A MEDLINE abstract and its TRSM representation

3. EXTRACTION OF ASSOCIATIONS OF TERMS FROM MEDLINE WITH TRSM-BASED SURROGATES AND APRIORI

The purpose of this work is to propose a method of extracting background knowledge from large resources of documents, in particular the extraction of medical background knowledge from large medical literature such as MEDLINE. In medical data mining, there are at least two reasons do carry out such a research:

- The background knowledge provided by domain experts is not always available.
- Medical literature resources such as MEDLINE contain a huge volume of research results and precious information that could be exploited, and that could not be all known by some individual domain experts.

We may think of various types of knowledge at different levels. For example, assume that we need to have information from the target text collection based on a specific template. In order to achieve the task, the information system has to recognize some sign words/phrase corresponding to the field indices of the template and detect solutions of words/phrase for them. In dealing with free text, essential requirements to the system include parsing, morphological analysis and other techniques as in [14], and those requirements make the text processing complex.

We prefer here a quite simple approach, as what we would like to extract from MEDLINE abstracts is plausible sets of combinations of medical examinations in terms of each problem posed to the hepatitis database. For difficulties caused by various synonyms of target terms, abstracts in short length, no requirement of complex analysis of text, or expects for semantically richness of combinations to extract, our key idea is to find strong associations among TRSM-based surrogates of abstracts. In this section, at first, the framework of the approach is introduced, then, the key idea of TRSM-based surrogate is described followed by tips.

3. 1. Framework

We design the procedure to obtain interesting rules from the hepatitis database by specifying LUPC constraints according to the combinations of medical examinations extracted from MEDLINE abstracts as follows.

(1) Collection of abstracts for each target problem set to the hepatitis database.
(2) Text preprocess and synonym solution for each set of abstracts by the stop word list and synonym lists.
(3) Tolerance class generation of each target terms.
(4) Surrogate generation for each set of abstracts
(5) Extract strong associations of terms by Apriori
(6) Mining rules from the hepatitis database by LUPC with specifying combinations of attributes obtained in (5) as constrains

3. 2. Surrogates Generation based on TRSM

According to the recommendation of medical experts, the set $\mathcal{A}$ of the following fifteen tests (terms) are the most important in investigating hepatitis and they are in our particular current focus.

$\mathcal{A}$ = {GOT, GPT, TTT, ZTT, T-CHO, CHE, ALB, TP, PLT, WBC, HGB, D-BIL, I-BIL, T-BIL, ICG-15}

We wish to find, by using MEDLINE, subsets of $\mathcal{A}$ whose elements are related to each other. However, a

crucial problem occurred when searching associations from a set of MEDLINE abstracts is the following. As MEDLINE abstracts are usually short while $\mathcal{A}$ is usually small, it happens that there are no such direct associations of terms in MEDLINE. Motivated by overcoming this obstacle, we are concerned with the problem formulated in the following steps:

(a) To approximately find a surrogate for each MEDLINE abstract, particularly by extending the part of interest in the MEDLINE abstract.
(b) To find associations of terms from the set of surrogates of MEDLINE abstracts.

Figure 6 describes the algorithm to generate surrogates of abstracts from MEDLINE.

Input A set $\mathcal{D}$ of MEDLINE abstracts, a set $\mathcal{A}$ of attributes under consideration, and the list syn(t_i) of synonyms for each term t_i in $\mathcal{A}$.

Output A set $\mathcal{D}^*$ containing of surrogates of abstracts in $\mathcal{D}$.

1. For each abstract d_j in $\mathcal{D}$, converting into $\overline{d}_j$ by the following rule: if a term w occurring in d_j is a synonym of some term ti of A, replace w in d_j with t_i. Denote $\overline{\mathcal{D}}$ by the set of all $\overline{d}_j$.
2. For each t_i in $\mathcal{A}$, find its tolerance class $I_\theta(t_i)$ in $\overline{\mathcal{D}}$ with a given θ.
3. For each t_i in $\mathcal{A}$, if t_i in $\mathcal{A}$ and t_i i in $\overline{\mathcal{D}}$, then add all the terms in $I_\theta(t_i)$ to $\overline{\mathcal{D}}$, i.e.,

   $\overline{d}_j \leftarrow \overline{d}_j \cup I_\theta(t_i)$
4. Remove each term w in $\overline{d}_j$ if w is neither a t_i in $\mathcal{A}$ nor in with some t_i in $\mathcal{A}$. Output $\overline{\mathcal{D}}$ as $\mathcal{D}^*$.

Figure 6: Algorithm to find TRSM surrogates of document abstracts

The followings are elements used in describing the method:

- A subset $\mathcal{D}$ of abstracts in MEDLINE relating to the research target. For example, if the target is to investigate the effectiveness of interferon in the hepatitis treatment, the two terms "hepatitis" and "interferon" can be used to selecting a subset $\mathcal{D}$ of MEDLINE abstracts. For examples in this case $\mathcal{D}$ contains 8,264 abstracts while MEDLINE contains 109,850 abstracts related to hepatitis.
- A set $\mathcal{A}$ of terms of some particular interest (hereafter called interested terms). For example, concerning the hepatitis database our interested terms are fifteen tests in the above set $\mathcal{A}$.
- A synonym list for each term t_i in $\mathcal{A}$, denoted by syn(t_i), that is assumed to be given or to be determined. The determination of synonym lists can be viewed as a preprocessing task that aims to convert combined terms into one term. For example, "GOT" is often described in different expressions, e.g., "AST" as another name, or full names with multiple terms, like "aspartate transaminase", "glutamate oxaloacetate transaminase" or "glutamic-oxaloacetic transaminase", and so on. In order to unify such equivalent words into each of target terms, we prepare the synonym list of terms containing abbreviations and different medical names standing for their concepts. For example, syn(IFN) = {interferon, IFN}.
- A set $\mathcal{D}^*$ containing of surrogates of abstracts in $\mathcal{D}$.

3. 3. Other tips

Text Collections

Two test sets of text abstracts are collected from MEDLINE through the PubMed interface [1]. One is retrieved with specifying keywords "hepatitis" and "interferon" (the IFN set), the other is collected with "hbv" and "hcv" (the B&C set). Each of key word combinations is related to each of the specified goals for the hepatitis dataset. The numbers of retrieved abstracts in IFN set and B&C set via PubMed are 6,240 and 2,464, respectively without empty texts

.

Preparation of Synonym Lists for Interested Terms

A tolerance class of a term is generated according to the number of its co-occurrences with other terms, while one concept might be expressed with a set of terms. In our case, most of 15 interested terms are abbreviations of term combinations, which we need to focus on, but make the generation of tolerance classes unstable. To fix the variations of interested terms, we decide to convert such variations into the unified terms, and define the correspondence between interested terms and their variations in a synonym list. For example:

syn(T-BIL) = {T-BIL, total bilirubin, TBIL,},
syn(GPT)={GPT, ALT, alanineaminotransaminase, glutamate pyruvic transaminase, …}

Association Mining from Surrogates

By replacing abstracts by their surrogates, each abstract is represented by a list of interested terms, to which association mining technique for transaction data can be applied. We simply use Apriori algorithm implemented in Clementine by SPSS [15].

4. EXPERIMENTS ON MEDLINE ABSTRACTS

4. 1. Framework of Experiments

Principally, what to evaluate here should be how effective this method can allow to obtain sets of rules which are interesting for users. This can be checked by these two criteria: a) the rules obtained from the hepatitis with applying associations from MEDLINE abstracts are evaluated higher than rules obtained without association information; b) whether there is any evidence concerning to the associations of examinations from MEDLINE abstracts.

However, it is too difficult to directly evaluate the criterion a), as the mining results are affected by many factors besides the association information. Therefore, our preliminary evaluation on this method focuses on the goodness of the associations by checking whether the medical evidences for the associations exist or not, then, see the difference between rules obtained from the hepatitis TA data with/without association information of medical examinations.

Before finding strong associations of terms, we also would like to investigate the most appropriate threshold to generate tolerance classes of terms. The size of tolerance classes of target terms can be a measure to evaluate which threshold to select, while there is no clear measure for such a task.

4. 2. Tolerance Class Generation

Tolerance classes $I_\theta(t_i)$ of each term t_i in the abstract set vary according to the threshold θ. Table 1 shows the size change of tolerance class for each interested term in the IFN set with varying co-occurrence threshold θ from 2 to 100 (the column "theta" stands for θ).

As seen in the table, the sizes of tolerance classes of "GOT" and "GPT" are extremely larger than that of other terms, as same in B&C set. It seems natural because those examinations are typical indicators to represent the liver condition, and tolerance classes are determined by co-occurrence frequencies of term pairs.

Regarding tolerance classes of other interested terms in the IFN set, "albumin (ALB)", "platelet (PLT)", "bilirubin (BIL)" are relatively bigger than others, followed by "hemoglobin (HGB)" and "total bilirubin (T-BIL)", while "direct bilirubin (D-BIL)", "indirect bilirubin (I-BIL)" do not appear in the set. In the B&C set, the tolerance class of "thymol turbidity (TTT)" does not exist. The sizes of PLT, ALB, D-BIL, HGB, T-BIL and "white blood cell (WBC)" are larger. For lower threshold, tolerance classes of "total protein (TP)", "in-docyanine green test (ICG)"and "cholinesterase (CHE)" could have more than one member, while D-BIL, I-BIL and "total cholesterol (T-CHO)" always have tolerance classes with single member, themselves.

Table 1: Size of tolerance classes of interested terms in the "IFN set" at each threshold θ

theta	kwd tcho	kwd che	kwd alb	kwd tp	kwd plt	kwd wbc	kwd hgb	kwd bil	kwd dbil	kwd ibil	kwd tbil	kwd icg15	kwd got	kwd gpt	kwd ttt	kwd ztt
2	1	33	582	30	687	128	313	551	1	1	190	19	1100	3440	0	42
3	1	14	353	10	424	64	187	315	1	1	96	3	692	2522	0	20
4	1	5	245	2	307	38	125	218	1	1	59	1	524	2062	0	7
5	1	1	190	1	228	20	91	166	1	1	30	1	417	1760	0	1
6	1	1	154	1	177	14	74	128	1	1	20	1	341	1521	0	1
7	1	1	116	1	146	9	56	104	1	1	16	1	285	1368	0	1
8	1	1	97	1	123	8	42	82	1	1	11	1	249	1244	0	1
9	1	1	88	1	108	6	36	66	1	1	7	1	216	1152	0	1
10	1	1	74	1	91	4	30	55	1	1	5	1	189	1086	0	1
20	1	1	19	1	31	1	8	19	1	1	1	1	83	637	0	1
30	1	1	11	1	17	1	2	9	1	1	1	1	44	446	0	1
40	1	1	6	1	8	1	1	3	1	1	1	1	26	346	0	1
50	1	1	2	1	7	1	1	2	1	1	1	1	16	284	0	1
60	1	1	1	1	6	1	1	1	1	1	1	1	15	239	0	1
70	1	1	1	1	3	1	1	1	1	1	1	1	12	211	0	1
80	1	1	1	1	2	1	1	1	1	1	1	1	12	184	0	1
90	1	1	1	1	1	1	1	1	1	1	1	1	10	164	0	1
100	1	1	1	1	1	1	1	1	1	1	1	1	8	145	0	1

In this situation, to choose a threshold θ commonly for all interested terms is not appropriate. The alternate way of determining tolerance class for each interested term is to modify the tolerance classes for each term with different threshold θs and the members included in tolerance classes of other terms. Finally, the member terms determination of tolerance class of each interested term is designed as: 1) select θ to produce tolerance class of a target term with less than the maximum condition of members; 2) remove any member term in a tolerance class if the member term appears in to many other classes, for example, 1/3 of $\mathcal{D}$, or 15 classes.

With considering those above, we partially modify our definition for document approximation, in which specified terms, in this case, they are "GOT" and "GPT", are ignored to be replaced by their tolerance class. After replacement of interested terms by their tolerance classes, each abstract in the set is represented by its surrogate, in which all terms belong to I_θ (t_i) except "GOT" and "GPT"

4. 3. Finding Associations among words

The following is a case with minimum support = 0.1 and minimum confidence = 0.75. It also shows few variations of relation among terms.

Here shows some association rules found on the B&C set with $\theta = 8$:

alb(928) <= bil(1031) & hgb(1598) & plt(847)
(7:7.5%, 1.0)

alb(928) <= bil(1031) & hgb(1598) (7:7.5%, 1.0)
alb(928) <= bil(1031) & plt(847) (7:7.5%, 1.0)
alb(928) <= bil(1031) (45:48.4%, 1.0)
bil(1031) <= alb(928) & hgb(1598) & plt(847) (7:7.5%, 1.0)
bil(1031) <= alb(928) (45:48.4%, 1.0)
hgb(1598) <= alb(928) & bil(1031) & plt(847) (7:7.5%, 1.0)
hgb(1598) <= alb(928) & plt(847) (7:7.5%, 1.0)
hgb(1598) <= bil(1031) & plt(847) (7:7.5%, 1.0)
hgb(1598) <= plt(847) (38:40.9%, 1.0)
plt(847) <= alb(928) & bil(1031) & hgb(1598) (7:7.5%, 1.0)
plt(847) <= hgb(1598) (38:40.9%, 1.0)

By checking various associations, we found, for example, two pairs of tests (ALB, BIL) and (PLT, HGB) have strong relation among pairs of attributes from the total of 13 attributes (I-BIL, D-BIL, T-BIL are group in BIL).

According to Medical Encyclopedia provided by Medline Plus [16], the part of "bilirubin" definition says "*Bilirubin* metabolism begins with the breakdown of red blood cells. Red blood cells contain hemoglobin, which is broken down to heme and globin. Heme is converted to *bilirubin*, which is then carried by *albumin* in the blood to the *liver*". It means that ALB concerns with the BIL metabolism in the body and this pair of examinations is supported by medical evidence. This combination of examinations is used for identifying troubles at liver among organs, as also according to the definitions in [16], ALB test helps in determining if a patient has liver disease or kidney disease, or if not enough protein is being absorbed by the body, while BIL test is useful in determining if a patient has liver disease or a blocked bile duct. However, ALB and BIL are also in context of Jaundice, yellow skin, and of course, hepatitis, as well. This suggests that extension of interested terms may lead to more informative results. Similarly, the combination of PLT and HGB is also reasonable, as these values show the basic blood conditions.

We also check which combinations to be found by ignoring terms in such strong associations, i.e., ALB, BIL, HGB and PLT. The detected associations are of TBIL and combinations among CHE, TP, WBC or ICG15. These are candidates for LUPC's constraints.

4. 4. Mining Hepatitis database by LUPC

Here we report the difference among the sets of rules learned by LUPC with or without specifying inclusive constrains.

By using the query view of D2MS [17], we can view the class distribution in terms of a combination of ALB and DBIL, and that this combination can separate HBV from HCV as in Figure 7, where 13 on 184 HBV patients and 1 on 272 HCV patients belong to the rule "if both ALB and ZTT are normal and D-BIL fluctuates between Normal and High, then HBV".

From TA data for HBV&HCV, the condition with min_acc = 85%, min_cov = 10, 22, 4 and 1 rules are obtained for cases with no constraints, with inclusive constraints (ALB, BIL, CHE, TP) and another inclusive constraints (ALB, BIL), respectively. Almost all rules with constraints setting are included in the rules obtained without constraints. The only one rule by constraints (ALB, BIL) is "if D-BIL is normal and T-BIL fluctuates between Normal and High, then HCV" with 20 correct cases among 22 relevant cases, and it does not show any relation of BIL with ALB. It seems unavoidable because 408 on 456 patients have "N" for ALB. By relaxing min_acc = 75% and min_cov = 5, the number of rules increases to 7, in which, the rules contains ALB for the condition part always have value "N" for ALB. However, there are a few rules, in which combinations of examinations in the condition parts are similar to the associations obtained from MEDLINE surrogates, e.g., "If CHE =N and T-BIL =N/H and TP = N, then HCV".

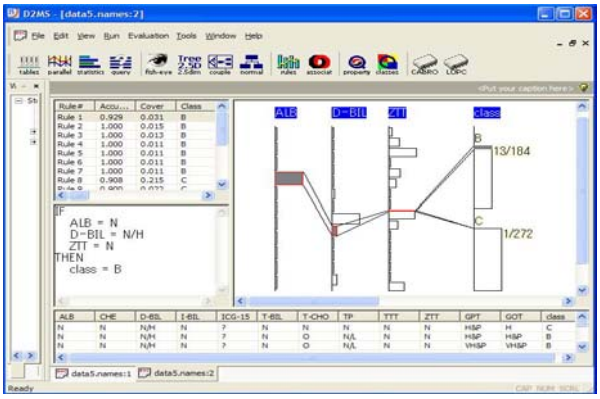

Figure 7: A rule found concerning ALB and DBIL

Meanwhile, TA data on IFN includes 128 cases and 102 cases belong to the class "response", 13 to "no response", 11 to "partial response" and 2 to "aggravation". From this dataset, LUPC always produce only four rules despite changing inclusive constraints for the conditions of min_acc = 75% and min_cov =5, For any of constraints set, the learned rule on IFN are same, in which either of ALB, T-BIL or D-BIL appears as only one condition of rules for the class "response". This is also resulted by the highly imbalanced class distribution.

5. CONCLUSIONS

We proposed a method of association extraction as background knowledge from MEDLINE by applying TRSM-based surrogate. Components of this method consist of: 1) framework of association extraction with TRSM-based surrogates and Apriori; 2) criteria for determining tolerance classes for converting abstracts to their surrogates. The experimental results suggest that this method has potentials to extract reasonable associations of medical examinations from MEDLINE supported by medical evidences. However, the rules obtained from the hepatitis data with specifying the association information from MEDLINE are too small to evaluate whether this framework contributes to produce interesting rules. Also, LUPC allows to specify constraints not only to attributes but also their values, so that it is useful to investigate methods to detect the associations of examinations with their values, that are more significant than the others.

REFERENCES

[1] NCBI http://www.ncbi.nlm.nih.gov/

[2] Cohen, A.M. and Hersh , W. R., "A Survey of Current Work in Biomedical Text Mining", *Briefings in Bioinformatics,* Vol. 6, No. 1, pp. 57-71, 2005.

[3] Hirschman, L., Park, J. C., Tsujii, J., Wong, L., and Wu, C. H., "Accomplishments and challenges in literature data mining for biology". *Bioinformatics,* Vol. 18, No.12, pp. 1553-1561, 2002.

[4] Ho, T.B., Nguyen, D.D., and Kawasaki, S., "Learning Minority Classes in Unbalanced Datasets", *Third International Conference on Parallel and Distributed Computing PDCAT 02*, Kanazawa, September 3-6, pp. 196-203, 2002.

[5] http://lisp.vse.cz/challenge/ecmlpkdd2004/HEP04-description.htm.

[6] Agrawal, R., Imielinski T., and Swami, A. N., "Mining Association Rules between Sets of Items in Large Databases", *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pp. 207-216, 1993.

[7] H. Motoda, "Analysis report on common data", *Active Mining*, Scientific Research on Priority Areas funded by MEXT, Japan, during 2001-2005, 2005.

[8] Shahar, Y. and Musen, M.A., "Knowledge-Based Temporal Abstraction in Clinical Domains", *Artificial Intelligence in Medicine*, Vol. 8, pp.267-298, 1996.

[9] Larizza, C., Bellazzi, R. and Riva. A., "Temporal abstractions for diabetic patients management", *Artificial Intelligence in Medicine*, Keravnou, E. et al. (eds.), Proc.AIME-97, pp 319-330, Springer, 1997.

[10] 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*, Washington DC, 24-27 August, 2003.

[11] Pawlak, Z.,"Rough sets: Theoretical Aspects of Reasoning about Data", Kluwer Academic Publishers, 1991.

[12] Polkowski, L. and Skowron, "A., Rough Sets in Knowledge Discovery: Applications", *Case Studies and Software Systems* (Eds.), Physica-Verlag, 1998.

[13] Ho, T. B. and Funakoshi K., "Information retrieval using rough sets", *Journal of Japanese Society for Artificial Intelligence*, Vol. 13, N. 3, pp. 424-433, 1998.

[14] Applet, D. E., and Israel, D. J. "Introduction to Information Extraction Technology", *tutorial at IJCAI-99*, 1999.

[15] Clementine http://www.spss.com/clementine/

[16] Medical Encyclopedia by Medline Plus http://www.nlm.nih.gov/medlineplus/encyclopedia.html

[17] Ho, T.B., Nguyen, T.D.,Nguyen, D.D., and Kawasaki, S., "Visualization Support for User-Centered Model Selection in Knowledge Discovery and Data Mining", *International Journal of Artificial Intelligence Tools, World Scientific*, Vol. 10, No. 4, pp.691-713, 2001.