DISCOVERY OF THE D-BASIS IN BINARY TABLES BASED ON
HYPERGRAPH DUALIZATION

K. ADARICHEVA AND J. B. NATION

Abstract. Discovery of (strong) association rules, or implications, is an im-
portant task in data management, and it finds application in artificial intel-
ligence, data mining and the semantic web. We introduce a novel approach
for the discovery of a specific set of implications, called the D-basis, that pro-
vides a representation for a reduced binary table, based on the structure of its
Galois lattice. At the core of the method is the D-relation defined in the lat-
tice theory framework, and the hypergraph dualization algorithm that allows
us to effectively produce the set of transversals for a given Sperner hyper-
graph. The latter algorithm, first developed by specialists from Rutgers Center
for Operations Research, has already found numerous applications in solving
the optimization problems in data base theory, artificial intelligence and game
theory. One application of the method is in analysis of gene expression data
related to a particular phenotypic variable, and some initial testing is done for
the data provided by the University of Hawaii Cancer Center.

1. Introduction

Knowledge retrieval from large data sets remains an essential problem in infor-
mation technology and its many usages in finance, biology, economy and social
sciences. The data is often recorded in binary tables with rows consisting of the
objects and columns of the attributes, that mark whether a particular object has or
does not have a particular attribute. The dependencies existing between the sub-
sets of the attributes in the form of association rules, or implications, can uncover
the laws, causalities and trends hidden in the data; see R. Agrawal et al. [3].

An implication $X \rightarrow Y$ (also referred to as a strong association rule in data
mining, see M. Kryszkiewicz [22], with both parameters of support and confidence
equal to 1) has the meaning that every object in the data set that possesses each
attribute in the subset $X$ will also possess attributes from the set $Y$. Such implica-
tions provide an essential hidden connection between different attributes. Sets of
implications which are called bases provide an alternative way of storing the tabled
data: they generate all possible implications holding in the set of attributes, as their
logical consequences, and also allow the restoration of the tabled data. Sometimes,
only the implications $X \rightarrow b$ pertinent to a distinguished attribute $b$ could be of
particular importance in various experiments that produce tabled data.

One particular type of a table could be medical data where $b$ may stand for
a phenotypic attribute, while other attributes are the levels of specific genes ex-
pressions. Then each association rule may represent a specific hypothesis of how

2010 Mathematics Subject Classification. 05A05, 06B99, 52B05.
Key words and phrases. Binary table, Galois lattice, implicational basis, hypergraph dualiza-
tion, association rules.

The research was partially supported by Grant N 13/42 of Nazarbayev University, 2013–2105.
the expression of several genes directly impacts the expression of another gene, and via a chain of direct interactions, indirectly impacts the phenotypic variation. Biologists may then pick some of these hypotheses to verify the interactions with the inventory of biochemical methods or animal models. To discriminate between the rules that are biologically relevant and those that are not, the concept of the support of the rule $X \rightarrow b$ may be applied, which is a portion of the rows of the table where all attributes from $X$ are present. The rules with the highest support could be the targets of prime interest.

In this paper we propose a novel algorithm for the retrieval of the strong association rules between the sets of attributes in large binary tables, which is particularly suited for retrieval of targeted implications $X \rightarrow b$ with a fixed attribute $b$. At the heart of the algorithm lies the connection between an arbitrary binary table and an associated algebraic structure known as a Galois lattice. A generating set of strong association rules, called a basis, often gives a more concise representation of the tabled data and its Galois lattice. Nevertheless, the existing algorithms developed in the framework of Formal Concept Analysis (FCA) are time-exponential on the size of the table, and they become ineffective on large data sets.

Our approach retrieves a complete generating set of implications known as the $D$-basis, that was introduced and tested in K. Adaricheva et al. [1]. This new type of basis is amenable to parallel processing, and can be used to find the rules that target a particular attribute without constructing the entire Galois lattice. Significantly, it employs the powerful optimization algorithm known as the hypergraph dualization, see M. Fredman and L. Khachiyan [15]. The code implementations of the algorithm were successfully tested on the models of the large size in E. Boros et al. [8]. While one can employ the algorithm for arbitrary tables, it could especially be effective in tables with the number of attributes considerably larger than the number of the objects. One finds this type of data in biological studies, where the set of attributes may include millions of genetic variations.

The paper is organized as follows. We present the association rules and parameters of support and the confidence, as they are known in data mining, in section 2. Then we give a general overview for the Galois lattices and FCA methods in section 3 and the hypergraph dualization algorithm in section 4. Section 5 describes the $D$-basis, and section 6 gives the general properties of Galois lattices. Our main result is in section 7 an algorithm to recover the $D$-basis from a binary table. We demonstrate the algorithm on an example in section 8 provide some preliminary results of testing the code implementation in section 9 and give an overview of the future work in section 10.

2. Association rules in data mining

The leading algorithmic approach in recovery of association rules in data mining was formulated in the form of the Apriori algorithm in R. Agrawal et al. [3]. Two parameters are essential for association rules: the thresholds of the support and the confidence. The support of any set $X$ of attributes is the value $S_A(X)$ (formally defined in section VI), which is essentially the number of rows of the table that have ones in all columns from $X$. The set $X$ is called $\delta$-frequent, if the ratio of $S_A(X)$ to the number of all rows in the table is greater than the given value of the threshold $\delta$. The first part of the Apriori algorithm aims at discovery of the maximal $\delta$-frequent sets, for a given threshold value $\delta$. 
The second part of the algorithm is devoted to splitting any maximal frequent set \( X \) into \( Y \cup Z \) so that the confidence \( S_A(Y \cup Z)/S_A(Y) \) exceeds a given threshold \( \mu \). In this case, \( Y \rightarrow Z \) is an association rule that satisfies both thresholds \( \delta \) and \( \mu \).

The time required to build maximal frequent sets, in the bottom-up way as in \textit{Apriori}, is asymptotically proportional to the number of frequent sets, and the latter may be exponentially larger than the number of maximal frequent sets. The number of maximal frequent sets itself may also be exponential in the size of the attributes of the table. The time required to build maximal frequent sets can be used as a norm to evaluate the complexity relative to the size of both input and output, in other words, the time delay in computation of the next maximal frequent set.

According to a result of E. Boros et al. [10], given any subset \( S \) of maximal frequent sets, the problem of deciding whether it is complete (i.e. includes all maximal sets) is NP-hard, even if the number of maximal frequent sets is exponential in the number of attributes \( n \), and \( S \) is of size \( O(n^\alpha) \) for small \( \alpha \).

All this makes the \textit{Apriori} heuristic inadequate for large data sets, in particular with many frequent sets of large size. Another weakness of the \textit{Apriori} approach is its inability to target specific attributes in the conclusion of association rules. Due to the nature of the algorithm, the decision about the splitting of a maximal frequent set into antecedent and conclusion happens in the second part of the algorithm, after the main effort on obtaining the frequent sets is already spent. See further details in section 9.

3. Formal Concept Analysis and basis retrieval

3.1. Retrieval of the canonical basis. The path from the tabled data to frequent sets of attributes, and then to bases of implications, can be built via the Galois relation existing between objects and attributes, and via the structure associated with this relation known in the algebraic literature as the Galois lattice. For example, maximal frequent subsets of the attribute set, which are targeted in data mining of transaction tables, are particular elements of the Galois lattice, see P. Valtchev et al. [28], while the strong association rules (with the support and confidence = 1) form the implicational basis of this lattice.

The first study of the Galois lattice appeared in 1940 in G. Birkhoff [7]. It was developed further in M. Barbut and B. Monjardet [5], and taken into a field of its own, under the name Formal Concept Analysis (or FCA), in B. Ganter and R. Wille [18]. FCA took off as a technical tool since the 1980s, mostly for the data visualization for various business applications. The applications range from knowledge representation and data mining to knowledge management and the semantic web, see J. Poelmans et al. [25].

Most recently, there were several projects in European Commonwealth where FCA was involved in biological studies of temporal Boolean networks modeling genetic data. See, for example, [2] [13] [29] [30]. There are growing applications of Galois lattices in ontology mappings [31] [32] [33] and in description logics [27].

The FCA approach is centered around a special implicational basis known in the literature as the canonical basis (also, stem basis, or Duquenne-Guigues basis). This basis has the minimum number of implications, and has a fundamental connection to any other basis defining a given Galois lattice and/or binary table.
The attribute exploration algorithm to recover the canonical basis, developed in the FCA framework, see B. Ganter [17], requires a search over the entire Galois lattice, with respect to some linear order established on the power set of attributes. As a result, the algorithm runs in times dependent on at least the size of the Galois lattice, which is normally exponential in the number of attributes. Recent theoretical results suggest that existing approaches to calculation of the canonical basis may not produce algorithms with better worst-case complexity, see [4, 12].

At its core, the implications of the canonical basis are defined by recognizing pseudo-closed sets $X$ that are placed into a premises of implications: $X \rightarrow Y$. There is no possibility to use parallel computation of pseudo-closed sets, since every new one depends on its subsets being recognized earlier in the process. Moreover, it may happen that the same attribute $b$ will appear on the right side in several implications from the basis. Thus, if one would be interested only in implications of the form $X \rightarrow b$, for this particular $b$, one would need to reconstruct the whole basis.

### 3.2. Canonical unit basis

Canonical unit basis is discussed in K. Bertet and B. Monjardet [6] as a basis unifying various definitions given to this concept in the literature. As the Duquenne-Guigues basis, it can be defined in the abstract framework of the general closure system on a finite set. It consists of implications $X \rightarrow b$ with the property of minimality of $|X|$, for any fixed $b$; in other words, no implication $X' \rightarrow b$ holds, where $X'$ is a proper subset of $X$.

The canonical unit basis has considerably more implications compared to the Duquenne-Guigues basis, but it has a nice feature of being iteration-free basis, or direct basis. In fact, it is contained in any other direct basis defining the same closure system.

Recently, U. Russel et al. [26] proposed a method of retrieval of the canonical unit basis that would employ the hypergraph dualization algorithm. While the actual coding applied some back-tracking and simple heuristics instead of actual implementation of dualization, the computer test results showed considerably shorter running times when compared to existing FCA implementations: on a particular data table of size $26 \times 79$ the new algorithm would return the basis of 86 implications in 0.1 sec compared to 6.5 hours of the FCA implementation, and, on random tables of size $20 \times 40$, it spent about 1/50 of time needed for the FCA algorithm.

### 4. Hypergraph dualization

Let $A$ be a finite set of cardinality $|A| = n$. For a hypergraph (set family) $H \subseteq 2^A$, let us denote by $I(H)$ the family of its maximal independent subsets, i.e., maximal subsets of $A$ not containing any edges of $H$. The complement of a maximal independent subset is a minimal transversal of $H$, i.e., a minimal subset of $A$ intersecting all edges of $H$. (Minimal transversals are also called minimal hitting sets.)

The collection $H^d$ of minimal transversals is called the dual or transversal hypergraph of $H$. It is easy to see that $H^d$ is a Sperner hypergraph, i.e., no edge of $H^d$ contains another edge of $H^d$. If $H$ is also Sperner then $H = (H^d)^d$. Given a Sperner hypergraph, a frequently arising task is the generation of the transversal hypergraph $H^d$. This problem, known as dualization, can be stated as follows:
DUAL($H, G$): Given a complete list of all edges of $H$, and a set of minimal transversals $G \subseteq H^d$, either prove that $G = H^d$, or find a new transversal $g \in H^d \setminus G$.

Clearly, one can generate all of the minimal transversals in $H^d$ (equivalently, all the maximal independent sets for $H$) by initializing $G = \emptyset$ and iteratively solving the above problem $|H^d| + 1$ times. Since $|H^d|$ can be exponentially large in both $|H|$ and $|A|$, the complexity of generating $H^d$ is customarily measured in the input and output sizes.

According to a result in \cite{14}, the problem DUAL($H, G$) can be solved in incremental quasi-polynomial time, i.e., in $O(n + m \log(m))$ time, where $n = |A|$ and $m = |H| + |X|$. Moreover, $H^d$ can be generated in incremental polynomial time (i.e. DUAL($H, X$) can be solved in time polynomial in $|A|, |H|$, and $|X|$) for many classes of hypergraphs, see \cite{8, 9, 14}.

The hypergraph dualization algorithm is one of various optimizations in database theory, artificial intelligence, game theory, and learning theory, to name a few. See the survey article \cite{8}. Its efficient implementation is described and tested in L. Khachiyan et al. \cite{21}. While it was shown that the implementation achieves the same theoretical worst bound, practical experience with this implementation shows that it can be substantially faster. In particular, the code can produce, in a few hours, millions of transversals for hypergraphs with hundreds of vertices and thousands of hyper-edges. Furthermore, the experiments also indicate that the delay per transversal scales almost linearly with the number of vertices and number of hyper-edges. A more recent implementation by K. Murakami and T. Uno \cite{24} not only demonstrates even better time performance, but also runs fast on large-scale inputs, for which earlier algorithms do not terminate in practical time.

5. D-Basis

The idea of the $D$-basis comes from concept of the $D$-relation developed in lattice theoretic framework. The definition of this relation go back to the work of B. Jónsson, A. Day, R. Freese, and J.B. Nation, which showed that this relation played a critical role in the description of the structure of free lattices. See the monograph R. Freese et al. \cite{16}. The $D$-relation plays a key role in defining the OD-graph of a finite lattice in J.B. Nation \cite{23}, which was widely used in computer science literature. This concept was translated into the $D$-basis in the recent work \cite{1}.

Essentially, the $D$-basis is a subset of the canonical unit basis that consists of implications $X \to b$ such that whenever any $x \in X$ is replaced by any set $Y$ for which $x \to y$ holds for all $y \in Y$, and $Y \to x$ does not hold, the implication no longer holds.

An important feature of the $D$-basis is that it is ordered direct, which means that it is iteration-free when a special ordering is imposed on the implications. This property of the basis has an advantage of easy parsing for the processing of its logical consequences. Such processing shows faster times than the forward chaining algorithm, which is widely used in industrial logic programming, or LINCLOSURE algorithm in data bases, see \cite{1}, section 7. Thus it retains the important property of directness of the canonical unit basis, while having, on average, only a portion of implications from the latter. According to the test results in \cite{1}, the average
number of implications in the $D$-basis, for the closure systems on domain of size 7, are between 0.55 and 0.70 of the size of the canonical unit basis.

6. Galois lattice

6.1. Support function and concepts. Many data sets in computer science are presented in the form of binary tables. By a finite binary table we understand a relation $R \subseteq U \times A$, for the set of objects $U$ (rows of the table) and set of attributes $A$ (columns of the table). If $r = (u, a) \in R$, then the position in row $u$ and column $a$ is marked by 1. This can be interpreted as meaning that object $u$ possesses attribute $a$. Otherwise, the position is marked with a 0.

This table represents a Galois connection, and allows us to form the corresponding Galois lattice. In order to define the Galois lattice, one needs to define the support function on each of the sets $2^U$ and $2^A$ with respect to $R$.

$s_A : 2^A \rightarrow 2^U$ is called a support function on $2^A$ if, for every $X \subseteq A$, $s_A(X) = \{ y \in U : (x, y) \in R, \text{ for all } x \in X \}$. Similarly, the support function $s_U : 2^U \rightarrow 2^A$ is defined for all $Y \subseteq U$ as $s_U(Y) = \{ x \in A : (x, y) \in R, \text{ for all } y \in Y \}$. One may use the symbol $S$ for notation of both $s_A$ and $s_U$, since it is often clear from the context which one should be applied.

The pair $(X, Y) \in 2^A \times 2^U$ is called a context of the relation $R$, if $Y = S(X)$ and $X = S(Y)$. It is easy to show that if $(X_1, Y_1)$ and $(X_2, Y_2)$ are two contexts with $X_1 \subseteq X_2$, then $Y_2 \subseteq Y_1$. Thus, the set of all contexts can be ordered with respect to the set containment order on their first components, or with respect to containment order of their second component, producing two partially ordered sets that are dual to each other.

Moreover, each of these partially ordered sets actually forms a lattice, $L_R$, or respectively, $L'_R$, and one may refer to one or the other (depending on the preferences between sets $U$ or $A$) as the Galois lattice (concept lattice in FCA) of the relation $R$.

Since the structure of the Galois lattice $L_R$ is fully determined by its first component (or its dual lattice $L'_R$ is fully determined by its second component), one can establish an isomorphism between the Galois lattice $L_R$ and the lattice of closed sets of a closure operator defined on $2^A$ by means of the support function. Indeed, it is straightforward to show that the operator $\phi_A : 2^A \rightarrow 2^A$ defined as $\phi_A(X) = S_U(s_A(X))$ for $X \in 2^A$ is, in fact, a closure operator on $A$, and the closed sets with respect to $\phi_A$ are exactly the first components of the concepts of $R$. Thus, $\text{Cl}(A, \phi_A) = L_R$, where $\text{Cl}(A, \phi_A)$ denotes the lattice of closed sets of a closure system $(A, \phi_A)$.

6.2. Reductions of the table. There exists a well-developed procedure of reducing the given relation $R$ to a relation $R' \subseteq U' \times A'$, where $U' \subseteq U$, $A' \subseteq A$ and $R' = R |_{U' \times A'}$, so that $|U'|$ and $|A'|$ are minimal with respect to property $L'_R \simeq L_R$. In other words, one may leave only essential objects and attributes in the relation and remove the others. Any analysis can be done on the smaller table, while easily expanding the outcome to the original sets of objects and attributes. In fact, every removed element $a \in A \setminus A'$ will be associated with some $X_a \subseteq A'$ such that $\phi_{A'}(a) = \phi_A(X_a)$.

When the procedure of reduction is completed and one obtains the reduced table $R'$, one can establish an important relationship between the elements of the set of objects $U'$, the set of attributes $A'$, and the special subset of elements in the Galois
lattice $L'_R$. Namely, the set $A'$ can be interpreted as the set of join irreducible elements $\text{Ji} L'_R$ of the Galois lattice $L'_R$, and $U'$ as the set of meet irreducible elements $\text{Mi} L'_R$. Moreover, an element 1 at the intersection of row $i$ and column $j$ is equivalent to $j \leq i$ in the lattice $L'_R$.

6.3. Arrow relations in the table. We may assume that after the reduction the table has rows of $n$ objects and columns of $m$ attributes. The binary table allows one to quickly recover additional information on $L'_R$:

1. Establishing a partial order $(U', \leq) = (\text{Mi} L'_R', \leq)$;
2. Establishing a partial order $(A', \leq) = (\text{Ji} L'_R')', \leq)$;
3. Establishing arrow relations $\uparrow$, $\downarrow$, and $\leftrightarrow$.

Recall that for $i \in \text{Ji} L'_R$ and $j \in \text{Mi} L'_R$, $i \uparrow j$ is defined to hold if $j$ is a maximal element among elements of $L'_R$ that are not greater than $i$. (In lattice terms, it is equivalent to: $i \lor j = j^*$, where $j^*$ is the unique upper cover of $j$).

Dually, $i \downarrow j$ if $i$ is a minimal element among elements of $L'_R$ that are not less than $j$. (This is equivalent to: $i \land j = i_*$, where $i_*$ is the unique lower cover of $i$). Finally, $i \leftrightarrow j$, if both $i \uparrow j$ and $i \downarrow j$ hold. It is clear that algorithmically, the reconstruction of arrow relations is equivalent to finding maximal or minimal elements in a particular partially ordered set, which are sub-posets of either $(U', \leq)$ or $(A', \leq)$. So we could apply appropriate algorithms to recover these relations. Thus, we may assume that the reduced table of lattice $L'_R$ is given equipped with the arrow relations.

6.4. Implicational basis. The Galois lattice can be fully determined by a set of implications defined on the set $A'$, or dually, on the set $U'$. By definition, an implication on set $A'$ is an ordered pair $(X, Y) \in 2^{A'} \times 2^{A'}$ with $X, Y \neq \emptyset$. Very often the implication $(X, Y)$ is written in the form $X \rightarrow Y$. A subset $Z \subseteq A'$ respects an implication $X \rightarrow Y$, if whenever $X \subseteq Z$ one also has $Y \subseteq Z$. If $\Sigma$ is a set of implications, then $Z$ respects $\Sigma$ whenever $Z$ respects every implication $\sigma$ from $\Sigma$.

There exists a classical connection between closure operators defined on $A'$ and sets of implications on $A'$:

1. every set of implications $\Sigma$ defines a closure operator $\phi_\Sigma$ by setting $\phi_\Sigma(Y)$ as the smallest overset of $Y$ that respects $\Sigma$;
2. every closure operator $\phi$ can be defined by some set of implications $\Sigma$ such that the $\phi$-closed sets are exactly sets that respect $\Sigma$.

While every set of implications defines the closure operator uniquely, there are multiple possibilities to define a set of implications for any given operator. The set of implications for a given closure operator, satisfying some conditions of minimality, is usually called an implicational basis of this closure system. One can refer to survey article [10] for further details on connections between closure operators, their closure lattices, and sets of implications.

For application purposes, most often the implications on $A'$ provide an essential hidden connection between different attributes in the given data set. For example, an implication $a_1 a_2 \rightarrow b$, where $a_1, a_2, b \in A'$, means that every object in the data set that possesses both attributes $a_1$ and $a_2$ also possesses attribute $b$. 


6.5. Connection between $D$-basis and $D$-relation. As we mentioned earlier, the name of the $D$-basis is directly connected to the $D$-relation defined in lattice theory framework. Lemma 2.31 in [16] can be formulated as follows.

Lemma 1. Given two elements $b, c \in J_i L$, the relation $bDc$ holds iff there exists an implication $X \rightarrow b$ in the $D$-basis of $L$ such that $c \in X$.

In particular, for every implication $X \rightarrow b$ in the $D$-basis, we have $X \subseteq bD = \{x \in J_i L : bDx\}$.

7. Recovery of the $D$-basis from the table

The goal of this section is to prove the main result of the paper.

Theorem 2. Given a reduced table $(U', A', R'), R' \subseteq U' \times A'$, one can polynomially (in the size of the table) reduce the problem of recovery of the $D$-basis to (parallel) solution of the hypergraph dualization problem formed for each $b \in A'$.

Proof. Recall from section 6.3 that set $A'$ can be interpreted as the set of join irreducible elements $J_i L$ of the Galois lattice $L$. The key observation for the recovery of the $D$-basis is contained in [16, Lemma 11.10] that connects the arrow relations of the reduced table with the $D$-relation on the set of join irreducible elements of its Galois lattice:

$$bDc \text{ iff } b \uparrow p \text{ and } c \downarrow p, \text{ for some } p \in Mi L_R.$$ 

This allows us to recover effectively sets $bD = \{c \in A' : bDc\}$, for every $b \in J_i L$.

According to the statement in section 6.5, for every fixed $b$, every implication of the $D$-basis of the form $X \rightarrow b$ will satisfy $X \subseteq bD = \{x \in J_i L : bDx\}$.

Another important subset associated with each $b$ is $M(b) = \{m \in Mi L : b \uparrow m\}$. Recall that these are the maximal elements in $L$ which are not greater than $b$. The following statement is a simple lattice theoretical observation.

Claim. For every $b \in J_i L$ and $X \subseteq J_i L$, $X \rightarrow b$ holds in $L$ iff for every $m \in M(b)$ there exists $x \in X$ such that $x \not\leq m$.

This observation allows us to reduce the problem of finding the $D$-basis to the problem of finding the minimal transversal sets of a particular Sperner hypergraph defined on the set $bD = \{c \in A' : bDc\}$. Namely, consider the hypergraph on $A'$ defined by the family of hyper-edges $H = \{H_m = bD \cap [0,m] : m \in M(b)\}$.

Here $0$ stands for the smallest element of lattice $L$ and $[0,m] = \{x \in L : x \leq m\}$. For each $m$, this is equivalent to taking only those elements from $bD$ that are in the relation $R'$ with $m$. The hypergraph dualization problem would search for the minimal elements of $2^{bD} \setminus J$, where $J$ is the order ideal in $2^{bD}$ generated by hyper-edges from $H$.

If $X_1, X_2, \ldots, X_k$ are such minimal elements, then the set of implications $X_i \rightarrow b$, $i = 1, \ldots, k$, gives a set of implications satisfying two properties:

- $X_i$ is a minimal subset $X$ such that $X \rightarrow b$ holds;
- $X_i \subseteq bD$.

In particular, all these implications belong to the canonical unit basis, and every implication from the $D$-basis of the form $X \rightarrow b$ is included into this list. The full basis will be recovered, when this process is applied to all $bD \neq \emptyset$, $b \in A'$.

As a result, the problem to recover the $D$-basis is polynomially reduced to at most $t$ runs of the dualization algorithm on $2^X$, $X_i \subseteq A'$, where $t = |A'|$ is the number
of the attributes. It is possible, however, that some of the recovered implications are not in the $D$-basis, so the recovered set of implications may contain the $D$-basis properly.

8. Example

Let us consider the procedure on a small table with 6 objects and 7 attributes. All symbols in the table other than 1 should be first interpreted as 0s.

|    | $b$ | $a_1$ | $a_2$ | $c_1$ | $c_2$ | $u$ | $v$ |
|----|-----|-------|-------|-------|-------|-----|-----|
| 1  | ↑   | ↑     | ↑     | 1     |       | 0   |     |
| 2  |     | ↑     |       |       |       |     |     |
| 3  | ↑   | ↑     |       | 0     | 0     | 0   |     |
| 4  |     |       |       | 1     | 1     | 0   |     |
| 5  |     |       |       | 1     | 1     | 1   | 0   |
| 6  |     | 1     | 1     | 1     | 1     | 1   | 1   |

Here $U = \{1, 2, 3, 4, 5, 6\}$, $A = \{b, a_1, a_2, c_1, c_2, u, v\}$. Since $S(c_1) = S(u)$, attribute $u$ can be reduced. Attribute $v$ also can be reduced, because $S(S(u)) = A$, and $A \setminus v$ is not a first component of any context. One can also reduce object 5 due to $S(5) = S(4)$, and reduce object 6 because $6 \in S(S(i))$, for every $i \in \{1, 2, 3, 4, 5\}$. Thus, one can consider the reduced table with $U' = \{1, 2, 3, 4\}$, $A' = \{b, a_1, a_2, c_1, c_2\}$, and complement the basis (on $A$) of a reduced table by implications $c_1 \rightarrow u$, $u \rightarrow c_1$, $v \rightarrow A \setminus v$.

The order relation on $A'$ consists of $c_1 \leq a_1$ and $c_2 \leq a_2$. The order relation on $U'$ consists of $2 \geq 4$. This allows us to determine the arrow relations between elements of $A'$ and $U'$, which are shown in the table. The Galois lattice $L'_R$ of the reduced table in shown on Fig. 1.

![Galois lattice from the table](image)

In order to recover all implications $X \rightarrow b$, for some $X \subseteq A'$, first identify elements of $bD = \{a_1, a_2, c_1, c_2\}$. Also, $M(b) = \{1, 3, 4\}$. Taking any $m \in M(b)$, we find a corresponding hyper-edge: $H_1 = \{a_1, c_1\}$, $H_3 = \{a_2, c_2\}$, $H_4 = \{c_1, c_2\}$. This
is done by picking support of each element $m \in M(b)$ within $bD$. The hypergraph dualization problem finds the minimal elements in $2^{bD} \setminus J$, where $J$ is the order ideal in $2^{bD}$ generated by hyper-edges $H_1, H_3, H_4$. Evidently, these will be $X_1 = \{a_1, c_2\}, X_2 = \{a_2, c_1\}$, and $X_3 = \{a_1, a_2\}$. This gives all implications from the basis with the conclusion $b$: $a_1c_2 \rightarrow b$, $a_2c_1 \rightarrow b$ and $a_1a_2 \rightarrow b$.

We note that the last implication is not in the $D$-basis and one can use the algorithm from [1] to remove it.

9. Results of initial testing

The algorithm of $D$-basis recovery described in section 7 was implemented in C++ code by the team of undergraduate students of Yeshiva University. They were able to implement the call to an existing subroutine that performs the hypergraph dualization, and which is publicly available via repository maintained by T. Uno: [http://research.nii.ac.jp/~uno/dualization.html](http://research.nii.ac.jp/~uno/dualization.html)

Further optimizations of the code were done in collaboration with T. Uno and U. Norbistrath. The retrieval of the full implicational basis for a random matrix $20 \times 40$ and density 0.2 (20% of entries of the matrix are ones) took 0.27 sec, for 1616 implications. Similar test in U. Russel et al. [26] mentioned in section 3 showed 0.9 sec for 1476 implications.

Further tests with the new code were done on larger random matrices. On the table of size 50-by-100, the $D$-basis with 49,000 implications was obtained in 3 min 30 sec. All the implications $X \rightarrow b$, for a requested column $b$ of randomly generated matrix of size 50-by-200, were obtained in 25 min.

For the initial comparisons with Apriori algorithm we ran two types of tests.

The first data set dataset was taken from the Frequent Itemset Mining Dataset Repository, publicly available at [http://fimi.ua.ac.be/data/retail.dat](http://fimi.ua.ac.be/data/retail.dat). It was retail market basket data from an anonymous Belgian retail store. We took first 90 rows converting them to a binary matrix format with size 90-by-502 and (low) density 0.0162. Running time was about 42 sec resulting in 104 mostly binary implications of maximal support 5 and confidence = 1. For the Apriori, we used Microsoft SQL Server Business Intelligence Development Studio, 2008 (Data Mining Technique - Microsoft Association Rules) [20][34]. The result of this run, together with set up of the input took about 4 min 30 sec, with considerably larger set of association rules, most of which have confidence < 1. We point that Apriori was designed specifically for mining the association rules in retail data, and the specifics of such data is that it presented by large sets of item-sets. When converted to matrix form, it usually contains large number of rows (transactions) with comparably low set of columns (number of items at sale), and normally has the low density.

Our second data set was of different nature, where we believe our novel approach may have an edge over Apriori.

The data-set was kindly provided by the research group of Dr. G. Okimoto from the University of Hawaii Cancer Center. It contained the gene expression levels for 550 pre-selected genes, in 22 patients, some healthy and others with liver cancer. There are multiple approaches how to convert this matrix into binary format. For comparison of $D$-basis algorithm with Apriori, one of them was taken as test data of size 22-by-1112 and density 0.4684. Columns 1111 and 1112 represent the attributes of being healthy and having cancer, correspondingly.
Microsoft Association Rules output was restricted to only 2000 frequent sets with minimum support 15 and confidence = 1, and there were no association rules with attributes 1111 and 1112. On the other hand, with the $D$-basis code we were able to reveal the equivalence of attribute 1111 to the set of 9 other attributes, and to find 14819 implications with the target attribute 1112, whose support = 8 and the confidence at least 21/22.

10. Future theoretical work and implementation

Recovery of the implicational basis directly relates to the knowledge discovery, especially if one is concerned with the targeted set of implications in the basis. The future testing of biological data provided by the collaborators at the University of Hawaii Cancer Center will be aiming at the data of the larger size, and at the retrieval of specific targeted association rules. The results of this testing will be presented in a separate publication.

From the theoretical point of view, the future plans include the generalization of the algorithm for the retrieval of the association rules with the threshold of confidence $\mu < 1$ (see section 2). This will allow us to take into account the possibility of incomplete and erroneous data. Theoretical results of the $D$-basis recovery presented in this paper are supported by the practical evidence tested in section 9 and initial testing presented in section 6 that the recovery of the implicational basis based on the hypergraph dualization algorithm will bring a considerable cut in run-times when dealing with tabled data. This will make the new approach critical for analysis of large data sets.

Acknowledgments. The results of this paper were prompted by the discussion of the hypergraph dualization algorithm with E. Boros and V. Gurvich, at the RUT-COR seminar during the first author’s visit in 2011. The test results of section 9 were possible due to code implementation done by J. Blumenkopf and T. Moldwin (Yeshiva College, New York), assistance of T. Uno (National Institute of Informat- ics, Tokyo), U. Norbisrath and A. Amantbekkyzy (Nazarbayev University, Astana), and data provided by G. Okimoto (University of Hawaii, Honolulu).

References

[1] K. Adaricheva, J.B. Nation and R. Rand, Ordered direct implicational basis of a finite closure system, Disc. Appl. Math. 161 (2013), 707–723.
[2] M. Agier and J-M. Petit, Defining, mining and reasoning on rules in tabular data, in Proc. 2nd Franco-Japanese Workshop on Information Search, Integration and Personalization, ISIP, Lyon, France, 2005.
[3] R. Agrawal, R. Sricant, H. Mannila, H. Toivonen and A. Verkamo, Fast discover of association rules, in Advances in Knowledge Discovery, U.Fayyad et al. eds., MIT Press, 1996.
[4] M.A. Babin and S.O. Kuznetsov, Computing premises of a minimal cover of functional dependencies is intractable, Disc. Appl. Math. 161 (2013), 742–749.
[5] M. Barbut and B. Monjardet, Ordres et classifications: Algebre et combinatoire, Hachette, Paris 1970.
[6] K. Bertet and B. Monjardet, The multiple facets of the canonical direct unit implicational basis, Theoretical Computer Science 411(2010), 2155–2166.
[7] G. Birkhoff, Lattice Theory, AMS Colloquium Publications 25 (1st ed), Providence, RI, 1940.
[8] E. Boros, K. Elbassioni, V. Gurvich and L. Khachiyan, Generating dual-bounded hypergraphs, Optimization Methods and Software 17 (2002), 749–781.
[9] E. Boros, V. Gurvich and P.L. Hammer, Dual subimplicants of positive Boolean functions, Optimization Methods and Software 10 (1998), 147–156.
[10] E. Boros, V. Gurvich, L. Khachiyan and K. Makino, *On the complexity of generating maximal frequent and minimal infrequent sets*, in Proceedings of STACS 2002, H. Alt and A. Ferreira eds., LNCS 2285 (2002), 131–141.

[11] N. Caspard and B. Monjardet, *The lattices of closure systems, closure operators, and implication systems on a finite set: a survey*, Disc. Appl. Math. **127** (2003), 241–269.

[12] F. Distel and B. Sertkaya, *On the complexity of enumerating the pseudo-intents*, Disc. Appl. Math. **159** (2011), 450–466.

[13] V. Duquenne, *Assessing implications between genotypic and phenotypic variables through lattice analysis*, Behavior genetics **31** (2001), 125–139.

[14] T. Eiter and G. Gottlob, *Identifying the minimal transversals of a hypergraph and related problems*, SIAM J. Comput. **24** (1995), 1278–1304.

[15] M. Fredman, and L. Khachiyan, *On the complexity of dualization of monotone disjunctive normal forms*, J. Algorithms **21** (1996), 618–628.

[16] R. Freese, J. Ježek and J. B. Nation, *Free Lattices*, Mathematical Surveys and Monographs **42**, Amer. Math. Soc., Providence, 1995.

[17] B. Ganter, *Attribute exploration with the background knowledge*, Theoretical Computer Science **217** N2 (1999), 215–233.

[18] B. Ganter and R. Wille, *Formal Concept Analysis*, Mathematical foundations, Springer Verlag, Berlin, 1999.

[19] J. L. Guigues and V. Duquenne, *Familles minimales d’implications informatives résultant d’une tables de données binaires*, Math. Sci. Hum. **95** (1986), 5–18.

[20] Jiawei Han, Micheline Kamber, *Data Mining Concepts and Techniques*, MK, 2006.

[21] L. Khachiyan, E. Boros, K. Elbassioni, and V. Gurvich, *An efficient implementation of a quasi-polynomial algorithm for generating hypergraph transversals and its application in joint generation*, Disc. Appl. Math. **154** (2006), 2350–2372.

[22] M. Kryszkiewicz, *Concise representation of frequent patterns and association rules*, Habilit. Thesis, Institute of Computer Science, Warsaw University of Technology, Poland, 2002.

[23] J.B. Nation, *An approach to lattice varieties of finite height*, Algebra Universalis **27** (1990), no. 4, 521–543.

[24] K. Murakami, and T. Uno, *Efficient algorithms for dualizing large scale hypergraphs*, Disc. Appl. Math. **170** (2014), 83–94.

[25] J. Poelmans, P. Elzinga, S. Vlaene and G. Dedene, *Formal concept analysis in knowledge discovery*, Proceedings of ICCS10 (2010), 139–153.

[26] U. Ryssel, F. Distel and D. Borchmann, *Fast algorithms for implication bases and attribute exploration using proper premises*, Ann. Math. Art. Intell. **70** (2014), 25–53.

[27] B. Sertkaya, *Formal concept analysis methods for description logics*, D.Sc. thesis, Dept. of Informatics, Technical University Dresden, Germany, 2008.

[28] P. Valttev, R. Massaoui, R. Godin, *A framework for incremental generation of closed itemsets*, Disc. Appl. Math. **156** (2008), 924–949.

[29] J. Wollbold, R. Huber, D. Pohlens, D. Koczan, R. Guthke, R. Kinne and U. Gausmann, *Adapted Boolean network models for extracellular matrix formation*, BMC Systems Biology, 2009.

[30] J. Wollbold, R. Guthke and B. Ganter, *Constructing a knowledge base for gene regulatory dynamics by formal concept analysis methods*, Algebraic Biology, Lecture Notes in Computer Science **5147** (2008), 230–244.

[31] Y. Yang, Y. Du, J. Sun and Y. Hai, *A topic-specific web-crawler with concept similarity context graph based on FCA*, Huang et al. (Eds.), in ICIC/LNAI **5227** (2008), Springer, 840–847.

[32] Y. Zhao, P.F. Shi, *Rough constrained lattice-based implication rules discovery*, The Journal of Shanghai Jiaotong University **35** N2 (2002), 177–180.

[33] Y. Zhao, Xia Wang and W. Halang, *Ontology mapping based on rough formal concept analysis*, Telecommunications, Proceedings of AICT/ICIW, Guadeloupe (2006), 180–187.

[34] Zhao Hui Tang and Jamie MacLennan, *Data Mining with SQL Server 2005*, Wiley, 2005.
Department of Mathematical Sciences, Yeshiva University, 245 Lexington Ave., New York, NY 10016, USA
E-mail address: adariche@yu.edu

Department of Mathematics, University of Hawaii, Honolulu, HI 96822, USA
E-mail address: jb@math.hawaii.edu