Scientific Knowledge Discovery through Iterative Transformation of Concept Lattices *

John L. Pfaltz
Christopher M. Taylor

Dept. of Computer Science, Univ. of Virginia
{jlp,cmt5n}@virginia.edu

April 8, 2002

1 Introduction

A basic goal of "data mining" is the discovery of frequent, or common, attribute associations in very large data sets such as point-of-sale data [1, 13, 14]. These associations are statistical in nature; for instance the old marketbasket example "people who buy hot dogs *often* buy catsup". They may also be evanescent. A frequent association this month may not be frequent next month. The association may be indicative of a causal relationship, but most often it is not. Surely the purchase of hot dogs does not "cause" a purchase of catsup. Frequent associations, whether temporary or not, can be very important to our understanding of the world around us. But, we would argue it is not for the most part a "scientific" understanding. Scientific knowledge primarily seeks the discovery and understanding of causal relationships among world phenomena that can be expressed with logical precision.

Let O denote any universe of interest. Our paradigm of scientific knowledge is an assertion of the form

$$(\forall o \in O)[P(o) \rightarrow Q(o)] \tag{1}$$

where $P(o)$ and $Q(o)$ denote predicate expressions over the bound variable o. Readily, there are other forms of scientific knowledge; but a focus on logical implication subsumes most causal assertions. In our development, we will regard O as a set of objects, or observations;

*Research supported in part by DOE grant DE-FG05-95ER25254.

we will treat predicates to be conjunctions and/or disjunctions of properties, or attributes, of these objects. A raw, unprocessed representation of this universe of objects is a binary relation $R : (O, A)$ whose rows correspond to objects, or observations, and whose columns correspond to attributes, which we collectively denote by A. Figure 1 is a typical example. It initiates a running example from which we will, in Section 2.2, eventually construct a

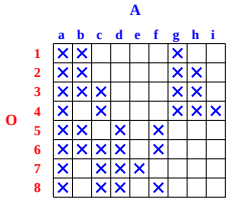

Figure 1: A small binary relation R from $O = \{12345678\}$ to $A = \{abcdefghi\}$

"concept lattice". The mathematics which we will first develop in Section 2.1 will then allow us to derive universally quantified expressions such as (1) above from these concept lattice structures. Finally, in Section 3, we show how to incrementally generate concept lattices, and their associated logical implications, as successive rows (scientific observations) are added to R. This kind of incremental discovery is not viable with many data mining methods because they require repeated sweeps over a fixed data set [9, 13]. This iterative form of knowledge discovery, which emulates rigorous scientific empiricism, is our major contribution.

2 Closure, Concepts and Implication

2.1 Closure Systems

We use the concept of closure to extract the desired logical implications from a relation, R. A closure operator φ is one that satisfies the three basic closure axioms: $X \subseteq X.\varphi$; $X \subseteq Y$ implies $X.\varphi \subseteq Y.\varphi$; and $X.\varphi.\varphi = X.\varphi$, for all X, Y.[1] There are many different closure operators. The geometrical convex hull operator is perhaps the most familiar [3], whereas monophonic closure on chordal graphs [4] is a bit obscure.

The intersection of any two closed sets (that is, those Z for which $Z.\varphi = Z$) of a closure space must also be a closed set of the space. The collection of all closed sets, partially ordered by inclusion, forms a lattice, $\mathcal{L}_\varphi$ [10, 11]. Of central importance to our development is the concept of "generators". A generator of a closed set Z is a minimal (w.r.t inclusion) set X such that $X.\varphi = Z$. By $Z.\gamma_i$ we mean the i^{th} generator of Z, and by $Z.\Gamma$ the collection $\{Z.\gamma_i\}$ of all generating sets.

[1] We use suffix notation to denote set valued operators. So read $X.\varphi$ as "X closure".

Figure 2 illustrates a closure lattice $\mathcal{L}_\varphi$ denoting a closure operator φ over a set, or

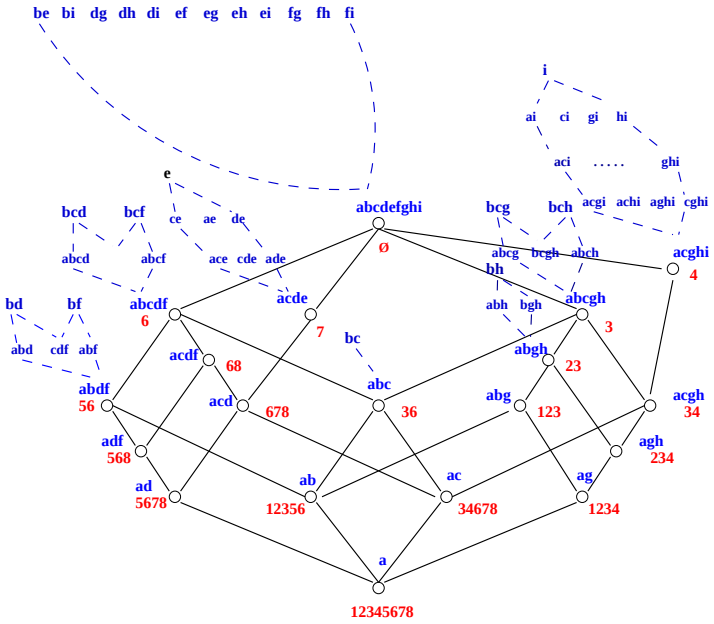

Figure 2: A lattice $\mathcal{L}_\varphi$ of closed sets with some generators shown.

universe $\mathbf{U} = \{a, b, c, d, e, f, g, h, i\}$ of elements. Solid lines connect the closed sets. By dashed lines we have tried to indicate a few of the generating sets. The singleton set $\{e\}$ is a minimal generator of $\{acde\}$.[2] The sets bd and bf are each minimal generators of $abdf$. Thus $abdf.\Gamma = \{abdf.\gamma_1, abdf.\gamma_2\} = \{bd, bf\}$.

If all generators are unique, the space is said to be antimatroid. Antimatroid closure spaces are particularly interesting [2, 6, 11]. But, readily, the closure space of Figure 2 is not antimatroid.

The whole space $abcdefghi$, is closed as required with any closure operator. It has 12 minimal generators, ranging from be through fi. The closure of any subset containing be, or any other generator, must be the whole set $\mathbf{U}$. We observe that, unlike closed sets which are closed under intersection, these generating sets are closed under union.

[2] From now on we elide the curly braces $\{\ldots\}$ around sets of elements of $\mathbf{U}$ whenever possible.

Let $\mathcal{F}$ be any family of sets. A set B is said to be a blocker for $\mathcal{F}$ if $\forall X \in \mathcal{F}, B \cap X \neq \emptyset$. The difference between a closed set Z and the closed sets Y_i that it covers in $\mathcal{L}_\varphi$ we call the faces F_i of Z.[3] Thus, the faces of *abcdf* are b, c and df. The faces of Z, its generators and blockers are closely related by

Theorem 2.1 *Let Z be closed and let $Z.\Gamma = \{Z.\gamma_i\}$ be its family of minimal generators.*
(a) If $X \subset Z$ and X is closed, then $Z - X$ is a blocker of $Z.\Gamma$.
(b) If B is a minimal blocker of $Z.\Gamma$, then $Z - B$ is closed.
(c) Z covers X in $\mathcal{L}_\varphi$ iff $Z - X$ is a minimal blocker of $Z.\Gamma$.

Proof:

(a) Let $Z.\gamma \in Z.\Gamma$ and suppose $Z.\gamma \cap (Z - X) = \emptyset$. Then, since $Z.\gamma \subseteq Z, Z.\gamma \subseteq X$. But, $Z.\gamma.\varphi = Z$ and thus $Z \subseteq Z.\gamma.\varphi \subseteq X.\varphi = X$, a contradiction.

(b) Let $Y = (Z - B).\varphi$. Then $Y \subseteq Z.\varphi = Z$. If $Y = Z$, then $Z - B$ is a generating set for Z, so it contains some minimal generating set $Z.\gamma$. Now, $Z.\gamma \subseteq Z - B$ implying $Z.\gamma \cap B = \emptyset$, contradicting assumption that B is a blocker of $Z.\Gamma$. So $Y \neq Z$.
Since Y is closed and $Y \subset Z$, by (a) $Z - Y$ is a blocker of Z. Because $Z - Y$ is a blocker, and because $Z - Y = Z - (Z - B).\varphi \subseteq Z - (Z - B) = B$, and because B is a minimal blocker, we have $B = Z - Y$. Thus $Y = Z - B$, and because Y is closed, $Z - B$ must be as well.

(c) readily follows from (a) and (b). If Z covers X in $\mathcal{L}_\varphi$, then $Z - X$ is a minimal blocker of $Z.\Gamma = \{Z.\gamma\}$; and if B is a minimal blocker of $Z.\Gamma$, then $X = Z - B$ is closed and Z covers X. □

This theorem can also be found in [**?**]. The closed set *abdf* and its generators *bd*, *bf* provide a good illustration of this theorem, since *bd* and *bf* are blockers of the two faces of *abdf*, $b = abdf - adf$ and $df = abdf - ab$ as asserted by the theorem.

2.2 Concept Lattices

Figure 2 was generated by applying a specific closure operator φ_R to the binary relation of Figure 1. By the closure, φ_R of O with respect to R, we mean a maximal set of objects which share the same attributes as all $o \in O$. Similarly, $\varphi_{R^{-1}}$ operating on a set A of attributes picks up any other attributes that are common to all objects which satisfy each $a \in A$. Ganter and Wille [5] show that φ_R and $\varphi_{R^{-1}}$ are indeed closure operators, and constitute a Galois connection.[4] A closure lattice, so generated is called a concept lattice.

[3] Recall that Z covers Y_i if $Y_i \subset Z$ and there exists no subset Y' such that $Y_i \subset Y' \subset Z$. The term "face" is derived from an application of closure in discrete geometry.

[4] More formally, the Galois closure, φ_R, on O with respect to R consists of those closed sets $\bar{O} \subseteq O$ of the form $\bar{O} = O_i.\overline{R}.\overline{R^{-1}}$, for $O_i \subseteq O$, where $O_i.\overline{R} = \bigcap_{o \in O_i} o.R \subseteq A$ and $A_i.\overline{R^{-1}} = \bigcap_{a \in A_i} a.R^{-1} \subseteq O$.

For any R, such as that of Figure 1, the closure systems of φ_R and $\varphi_{R^{-1}}$ are isomorphic and can be represented by the lattice $\mathcal{L}_R$ of closed sets shown in Figure 2. Labeling each node is the pair of closed sets that is joined by the Galois connection, for example $< abg, 123 >$. The set abg is closed in A; 123 is closed in O. In this case we have oriented the lattice with respect to A, the set of attributes, so that the universe $A = abcdefghi$ (which must be closed) is the lattice *supremum*. The singleton set $\{a\}$, which is an attribute of every object is the lattice $infimum$. It is partially ordered with respect to attribute set inclusion. The nodes of the concept lattice correspond to abstract *concepts* of the phenomenon being modeled and relationships within the lattice are reflective of relationships in the external world.

In Ganter and Wille's approach the visual display of concept lattices is crucial. For instance, the bifurcation into two distinct classes of concepts is visually evident in Figure 2. In their book [5] they give many clear, visual examples. But this approach becomes problematic with more than 30 concepts. One can no longer "see" the structure.

2.3 Implications

Concept lattices have appeared in the data mining literature where several authors [8, 14] have employed them to speed the search for frequent sets and their associations; but none have considered the role of generators.

Each closed concept of a concept lattice has generators, as discussed in Section 2.1. If we regard R as a relation in the database sense, then $(o, a) \in R$ implies that a is an attribute of object o. In Figure 1 it is clear that $abgh$ are shared attributes of objects 2 and 3. Attributes bh generate $abgh$. So we may assert that *in this world* $(\forall o \in O)[bh(o) \rightarrow abgh(o)]$; or more simply we have the attribute implication $bh \rightarrow abgh$ Similarly one may show that either bcd and bcf are minimal generators of $abcdf$; so we have the attribute implication $(\forall o \in O)[bcd(o) \vee bcf(o) \rightarrow abcdf(o)]$. We use $\rightarrow$ to denote both attribute implication and closure generation.

So far we have explored closure over attributes that are just sets of letters, which, by R, are associated with objects that are only integers. Such abstraction is the essence of mathematics, but it can lose immediacy in consequence. As reported in [5], the relation R of Figure 1 was actually derived from a child's educational TV program on pond life. The attributes are those discussed in the program to show differences between life forms. The rows represent observed life forms, or objects, which exhibit combinations of these

Conversely, one forms the closure, $\varphi_{R^{-1}}$ of A with respect to R consisting of the closed sets $\bar{A} = A_k.\overline{R^{-1}}.\overline{R}$. The set $O_i.\overline{R}$ denotes the set of all attributes shared by every object in O. Consequently, $O_i.\varphi = O_i.\overline{R}.\overline{R^{-1}}$ denotes the set of *all* the objects that share (at least) these common attributes. Similarly, $A_k.\overline{R^{-1}}$ denotes the set of all objects sharing every attribute in A_k and $A_k.\varphi = A_k.\overline{R^{-1}}.\overline{R}$ consists of *all* the attributes shared by the objects which (at least) have Y in common.

attributes. These semantics are listed in Figure 3.

Objects		Attributes	
1	leech	a	needs water to live
2	bream	b	lives in water
3	frog	c	lives on land
4	dog	d	needs chlorophyl to make food
5	spike-weed	e	two little leaves grow on germinating
6	reed	f	one little leaf grows on germinating
7	bean	g	can move about
8	maize	h	has limbs
		i	suckles its offspring

Figure 3: A semantic assignment to the rows and column headings of Figure 1.

Given these semantic assignments, the assertion $bh \rightarrow abgh$ becomes "if o lives in water and has limbs, then o can move about and needs water to live". And, $bcd \vee bcf \rightarrow abcdf$ can be interpreted as "for any o that both lives in water and lives on land, the properties 'needs chlorophyl to live' and 'one little leaf grows on germination' are equivalent". And, it is evident that the visual bifurcation noted earlier is the difference between the $flora$ and $fauna$ at the pond.

By deriving the generators of all closed concept sets, we extract all the logical implications (universally quantified over O) that are valid for R. Consider the eleven non-trivial generator-closure pairs: $e \rightarrow acde$, $f \rightarrow adf$, $h \rightarrow agh$, $i \rightarrow acghi$, $\{bd, bf\} \rightarrow abdf$, $\{be, bi, dg, dh, di, ef, eg, eh, ei, fg, fh, fi\} \rightarrow abcdefghi$, $bh \rightarrow abgh$, $cf \rightarrow acdf$, $\{cg, ch\} \rightarrow acgh$, $\{bcd, bcf\} \rightarrow abcdf$, $\{bcg, bch\} \rightarrow abcgh$. These fully describe all the inferable implications derivable in R. Consequently, Figure 2 can be regarded as a knowledge structure that completely represents a child's understanding of pond life, given these observations O of objects having attributes A.

Unfortunately, to calculate the generator-closure pairs using the definition at the beginning of Section 2.2 requires repeated looping over all objects $o \in O$. For large numbers of objects, or observations, this exponential process becomes prohibitive.

3 Incremental Growth

If a concept lattice $\mathcal{L}_R$ captures all the logical attribute implications one can make about a collection of objects; it is natural to ask "suppose we observe one more object and its attributes. How will this transform the lattice $\mathcal{L}_R$?" This is the essence of discrete, empirical induction. Our understanding of world phenomena seldom occurs in one fell swoop, as by examining a complete relation R. It normally occurs incrementally. We assimilate new observations, or rows of R if you choose, into an existing concept structure.

If we observe a new denizen of our pond life — a plant which lives in water (b), but has two little leaves on germination (e), that is abe, because all life (a) needs water to live, we can add a 9^{th} object (or row) having attributes a, b and e; the concept lattice of Figure 2 becomes that of Figure 4. The new relationships between concepts occasioned by abe are indicated by dotted lines.

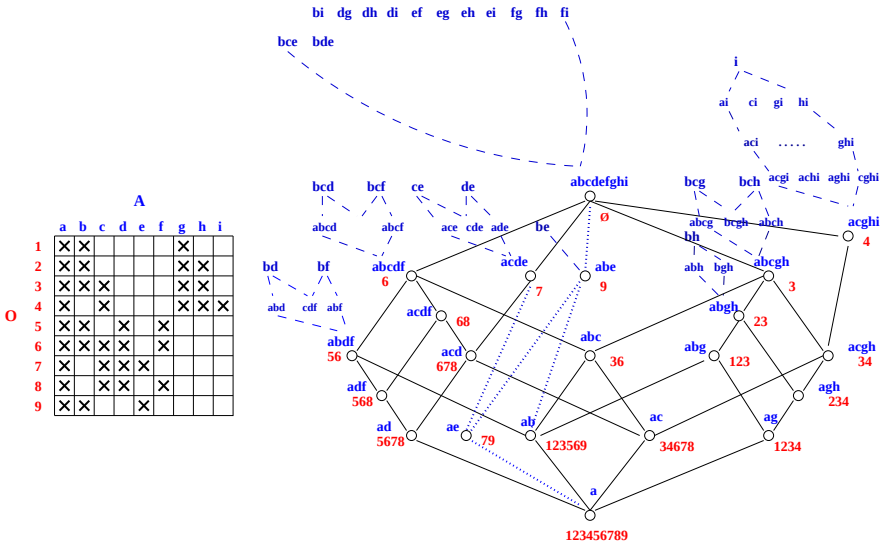

Figure 4: The concept lattice of Figure 2 after another object 9 has been incrementally added.

Readily, Figure 4 is a relatively smooth transformation of Figure 2. Closure properties ensure that observing new objects must lead to reasonably "graceful" transformations of these concept lattices [12]. Our goal is to make this notion more mathematically precise.

Let $f : \varphi \to \varphi'$ be closure preserving, that is Z closed w.r.t. φ implies Z is closed w.r.t φ'. But, there may be more sets closed with respect to φ'. We let $\mathcal{L}_\varphi$ denote the entire structure of a closure system induced by the operator φ.

Given $f : \varphi \to \varphi'$, by $f_Y : \mathcal{L}_\varphi \to \mathcal{L}'_{\varphi'}$ we denote the induced transformation of $\mathcal{L}_\varphi$ caused by regarding Y as closed with respect to φ'. Readily, if Y is already closed in $\mathcal{L}_\varphi$, then F_Y is simply the identity map which we call *trivial*.

Lemma 3.1 *Let $f_Y : \mathcal{L}_\varphi \to \mathcal{L}'_{\varphi'}$. For all closed $X \in \mathcal{L}_\varphi$ if $X \cap Y \neq \emptyset$, then $X' \cap Y'$ is closed in $\mathcal{L}'_{\varphi'}$.*

Proof: f_Y closure preserving implies $X.f_Y = X'$ is closed in $\mathcal{L}'_{\varphi'}$ and $Y = Y'$ is closed in $\mathcal{L}'_{\varphi'}$ by

definition of f_Y. The result follows from the basic property that all closure systems are closed under intersection. □

Godin and Missaoui [7] took advantage of this to incrementally create concept lattices in a way that minimizes the use of costly closure computation. But, their approach does not consider generators. We have extended this notion of incremental expansion to include the generating sets as well.

Lemma 3.2 *Let $f_Y : \mathcal{L}_\varphi \to \mathcal{L}'_{\varphi'}$ be non-trivial. There exists a single closed set Z in $\mathcal{L}$ such that Z' covers Y' in $\mathcal{L}'_{\varphi'}$.*

Proof: Suppose Y is covered by Z_1 and Z_2, both closed in $\mathcal{L}$, then $Y \subseteq Z_1 \cap Z_2 \subset Z_i$ which must be closed, hence in $\mathcal{L}$. □

Theorem 3.3 *Let $f_Y : \mathcal{L}_\varphi \to \mathcal{L}'_{\varphi'}$. Let Z' cover Y' in $\mathcal{L}'_{\varphi'}$, and let $Z.\gamma_i$ denote a generator of Z in $\mathcal{L}$.*

(a) If $Z.\gamma_i \cap (Z-Y) \neq \emptyset$, then $(Z.\gamma_i)'$ is a generator of Z' in $\mathcal{L}'$.

(b) If $Z.\gamma_i \cap (Z-Y) = \emptyset$, then for all $\alpha \in Z-Y$, $(Z.\gamma_i \cup \{\alpha\})'$ is a generator of Z', provided it is minimal.

Proof:

(a) If $Z_i \cap (Z-Y) \neq \emptyset$, then $(Z.\gamma_i)' = Z.\gamma_i.f_Y$ is still a generator of $Z' = Z.f_Y$, and it blocks the face $Z-Y$ (as required by Theorem 2.1). Minimality follows because otherwise we would have a smaller generating and/or blocking subset which would contradict the assumption that $Z.\gamma_i$ is a minimal generator.

(b) If $Z.\gamma_i \cap (Z-Y) = \emptyset$, then $Z.\gamma_i \subseteq Y$. And since $Z.\gamma_i.\varphi' \subseteq Y'$, $Z.\gamma_i$ cannot be a generator of Z' in $\mathcal{L}'_{\gamma'}$. Let $\alpha \in Z-Y$. $(Z.\gamma_i \cup \{\alpha\}).\gamma = Z$. So, since Z' covers Y' in $\mathcal{L}'_{\gamma'}$, $(Z.\gamma_i \cup \{\alpha\}).\gamma' = Z'$ and is hence a generator. $Z.\gamma_i \subseteq Y$. $Z.\gamma_i \cup \{\alpha\}$ is a blocker of every face. However, it need not be minimal, hence the caveat. □

The new closed set $Y = abe$ is covered by $Z = \mathbf{U} = abcdefghi$ in Figure 4 forcing $abcdefghi.\Gamma$ to change. Except for be, each generator $Z.\gamma_i$ has a non-empty intersection with the face $cdfghi = Z-abe$ and hence by Theorem 3.3(a) is a generator of Z' in $\mathcal{L}'_{\gamma'}$. But, $be = Z.\gamma_1 \subseteq abe = Y$, hence by part (b) of Theorem 3.3, be must be augmented by c, d, f, g, h, i in turn. But notice, bef is not a minimal generator since $ef \subseteq bef$. Only bce and bde are minimal, and they are shown in Figure 4.

By Lemma 3.1, if $Y = abe$ has a non-empty intersection with any other closed set in $\mathcal{L}$, then that set must also be in $\mathcal{L}'$. We need only check with the siblings of Y, that is other sets covered by $Z = abcdefghi$. These faces determine the generators $Y'.\Gamma$. For most, such

as $abcdf \cap abe = ab$, the intersection set is already in $\mathcal{L}'$. Only $X = acde \cap abe = ae$ yields a new set. This too must be recursively inserted into $\mathcal{L}'$ using the rules of Theorem 3.3 to change $acde.\Gamma$. In $\mathcal{L}$, $acde$ covered only acd, so it had a singleton generator e. In $\mathcal{L}'$, it covers both acd and ae, and so requires both ce and de to satisfy the requirements of Theorem 2.1

It is not hard to write a recursive procedure that captures the behavior of Theorem 3.3. With it we have found that iterative generation of concept lattices with generators is nearly an order of magnitude faster than the way suggested by the footnote in Section 2.2.

Suppose a child exploring the pond life sees a snake, that is a creature that also needs water(a), lives on the land(c), and can move about(e); but doesn't have limbs. We add a tenth observation acg to R. In a way, Figure 5 is much more representative of incremental growth. Most transformations f_Y tend to be very local in nature as is this one. Only

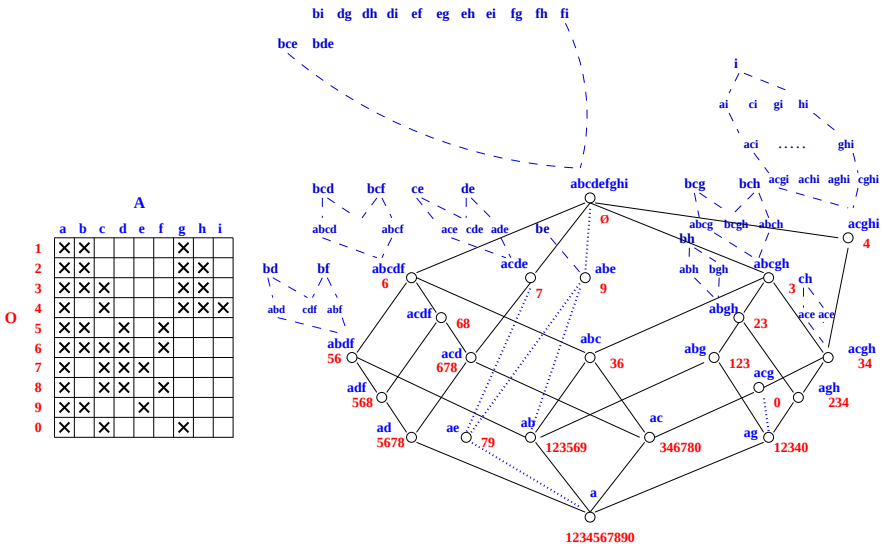

Figure 5: The concept lattice of Figure 4 after still another object 10 has been incrementally added.

$acgh.\Gamma$, which had been $\{cg, ch\}$ has changed in accordance with the rules of Theorem 3.3. Can you determine the generator(s) of acg given its faces with respect to ac and ag, which it covers, using Theorem 2.1?

The approach described in this paper offers an effective way of extracting all the valid, logical implications inherent in a large body of data. Moreover, these results indicate that the gradual accumulation of "knowledge" based on sequential, empirical observations inherent in data is relatively "stable". This is in accord with our intuitive understanding

of knowledge.

References

[1] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining Association Rules between Sets of Items in Large Databases . In *Proc. 1993 ACM SIGMOD Conf.*, pages 207–216, Washington, DC, May 1993.

[2] Brenda L. Dietrich. Matroids and Antimatroids — A Survey. *Discrete Mathematics*, 78:223–237, 1989.

[3] Paul H. Edelman and Robert E. Jamison. The Theory of Convex Geometries. *Geometriae Dedicata*, 19(3):247–270, Dec. 1985.

[4] Martin Farber and Robert E. Jamison. Convexity in Graphs and Hypergraphs. *SIAM J. Algebra and Discrete Methods*, 7(3):433–444, July 1986.

[5] Bernard Ganter and Rudolf Wille. *Formal Concept Analysis - Mathematical Foundations.* Springer Verlag, Heidelberg, 1999.

[6] Paul Glasserman and David D. Yao. Generalized semi-Markov Processes: Antimatroid Structure and Second-order Properties. *Math. Oper. Res.*, 17(2):444–469, 1992.

[7] Robert Godin and Rokia Missaoui. An incremental concept formation approach for learning from databases. In *Theoretical Comp. Sci.*, volume 133, pages 387–419, 1994.

[8] Robert Godin, Rokia Missaoui, and Hassan Alaoui. Incremental Concept Formation Algorithms Based on Galois (Concept) Lattices. *Computational Intelligence*, 11(2):246–267, 1995.

[9] Karam Gouda and Mohammed J. Zaki. Efficiently Mining Maximal Frequent Item Sets. In *1st IEEE Intern'l Conf. on Data Mining*, San Jose, CA, Nov. 2001.

[10] B. Monjardet. A Use for Frequently Rediscovering a Concept. *Order*, 1:415–416, 1985.

[11] John L. Pfaltz. Closure Lattices. *Discrete Mathematics*, 154:217–236, 1996.

[12] John L. Pfaltz. Transformations of Concept Graphs: An Approach to Empirical Induction . In *2nd International Workshop on Graph Transformation and Visual Modeling Techniques*, pages 320–326, Crete, Greece, July 2001. Satellite Workshop of ICALP 2001.

[13] Ramakrishnan Srikant and Rakesh Agrawal. Mining quantitative association rules in large relational tables. In *Proc. of 1996 SIGMOD Conference on Management of Data*, pages 1–12, Montreal, Quebec, June 1996.

[14] Mohammed J. Zaki. Generating Non-Redundant Association Rules. In *6th ACM SIGKDD Intern'l Conf. on Knowledge Discovery and Data Mining*, pages 34–43, Boston, MA, Aug. 2000.