A Bag-based Data Model with Global Conditions for Incomplete Information

Akinari YAMAGUCHI†, Shougo SHIMIZU††, Yasunori ISHIHARA†, and Toru FUJIWARA†

† Graduate School of Information Science and Technology, Osaka University
1–5 Yamadaoka, Suita, Osaka, 565–0871 Japan
†† Department of Computer and Media Technologies, Hiroshima City University
3–4–1 Ozuka-higashi, Asaminami-ku, Hiroshima, 731–3194 Japan

Abstract In our previous work, we proposed a data model, called V-bags, for representing incomplete information on bags of tuples. This paper proposes a supermodel of V-bags, called GV-bags, which are V-bags augmented by global conditions. Global conditions restrict assignments over variables in V-bags. Then, the closure properties of six operations, selection, projection, union, product, difference and unique, and their inverses on these models under both closed world assumption (called CWA) and open world assumption (called OWA) are investigated. We obtain that unique on V-bags is closed under CWA, but it under OWA and its inverse on V-bags is not closed under either CWA or OWA. It is shown that inverse of product under CWA, inverse of selection under OWA and inverse of unique under CWA are closed on GV-bags although they are not closed on V-bags. The others are the same as those on V-bags.

Key words incomplete information, algebraic operation, closure property, bag

1. Introduction

1.1 Background

Incomplete information is partial and ambiguous information such as "We know that the course Databases is given at room A01, but we *do not know* who teaches it" and "The room of the course Programming given by Ishihara is B02 *or* C03." Information which we obtain in the real world is often incomplete. To handle such incompleteness strictly, we need data models for representing incomplete information.

Moreover, in some of the recent works on answering queries using materialized views, data models for incomplete information are demonstrated to be useful [1], [6], [10], [13]. Such data models are also useful for access control guaranteeing no secret information disclosed. As an example, consider *Disclosure Monitor* proposed in [5]. When a database user issues a query, Disclosure Monitor maintains the following two types of knowledge: (a) knowledge which the user has obtained from the database so far, and (b) knowledge which the user can obtain from the query and the answer of the query. If the secret information of the database can be derived from knowledge (a) and (b), then Disclosure Monitor refuses to answer the query. Data models for incomplete information are useful for representing the user's knowledge because queries and their answers can be regarded as incomplete information on the database. Hence, if such incomplete information can be naturally treated, the idea of Disclosure Monitor will have more impact.

Ordinary relational databases cannot represent incomplete information naturally. Therefore, some extensions of relational databases have been proposed. For example, in a *Codd table* [7], unknown values are represented by one special symbol, *null*. In a *V-table* [11], unknown values are represented by variables. *C-tables* [11] are a supermodel of V-tables such that each tuple has a condition for the tuple to exist. In [3], the complexity of operations on those models is investigated.

Example 1 Figures 1 and 2 give examples of a V-table and a C-table, respectively. Figure 1 shows three facts that arbitrary teacher gives Databases at room A01, and Ishihara teaches two courses, Programming and Network. Figure 2 contains a new attribute *con* representing the condition for the associated tuple to exist. For example, the tuple (Programming, Ishihara, y) exists if y is B02 or C03.

The underlying data models of Codd tables, V-tables and C-tables are relational. However, most of the practical query languages contain constructs and functions which are represented on bags, i.e., the answer of a query may contain duplicate tuples. In SQL, the "select distinct" construct and the "count" function are examples of such constructs and functions. These operations can be better explained if bags instead of sets are used. Thus, a bag-based data model for representing incomplete information is desirable. Moreover, bag-based data models can be implemented easily on relational databases, by using a special attribute for representing the number of tuples. Also, they can be implemented on

Course	Teacher	Room
Databases	x	A01
Programming	Ishihara	y
Network	Ishihara	y

Figure 1　An example of a V-table.

Course	Teacher	Room	*con*
Databases	x	A01	true
Programming	Ishihara	y	$(y = \text{B02}) \vee (y = \text{C03})$
Network	Ishihara	y	$(y = \text{B02}) \vee (y = \text{C03})$

Figure 2　An example of a C-table.

Global condition: $(x \neq \text{Yamaguchi}) \wedge (z \leqq 3)$

Teacher	*Room*		
x	A01	$\mapsto$	z
Ishihara	y	$\mapsto$	"if $(y = B02) \vee (y = C03)$ then 2 else 0"

Figure 3　An example of a GV-bag.

Table 1　Closure properties.

(a) V-bags, CWA

	σ	π	$\cup$	$\times$	$-$	μ
Forward	Y	Y	Y	Y	Y	**Y**
Inverse	N	N	Y	N	N	**N**

(b) V-bags, OWA

	σ	π	$\cup$	$\times$	$-$	μ
Forward	N	Y	Y	Y	Y	**N**
Inverse	N	N	Y	Y	N	**N**

(c) GV-bags, CWA

	σ	π	$\cup$	$\times$	$-$	μ
Forward	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**
Inverse	**N**	**N**	**Y**	**Y**	**N**	**Y**

(d) GV-bags, OWA

	σ	π	$\cup$	$\times$	$-$	μ
Forward	**N**	**Y**	**Y**	**Y**	**Y**	**N**
Inverse	**Y**	**N**	**Y**	**Y**	**N**	**Y**

Y: the operation is closed.
N: the operation is not closed.

other models, i.e., there is a lot of flexibility in implementing bag-based data models.

1.2 Our results

In [17], we proposed a bag-based data model for incomplete information, called *EC-tables*, which has a special attribute of conditions for tuples to exist. In this paper, we propose a data model called *V-bags* for representing incomplete information on bags of tuples. V-bags do not have such a special attirbute of conditions but are equivalent to EC-tables bacause in V-bags, the number of tuples can be specified by a conditional expression. Like C-tables, incompleteness is represented by variables in V-bags. In [17], the closure properties of five operations, selection, projection, union, product and difference and their inverses are investigated. Since assignments over variables in V-bags cannot be restricted in this model, some operations are not closed. In this paper, we propose a supermodel of V-bags, called *GV-bags*, which are V-bags augmented by *global conditions* [4], [8], [9]. Global conditions restrict assignments over variables in V-bags.

Example 2　Figure 3 shows an example of a GV-bag. This GV-bag has a global condition $(x \neq \text{Yamaguchi}) \wedge (z \leqq 3)$. The numbers of tuples are shown to the right of $\mapsto$. Then, the first tuple describes that some teacher x other than Yamaguchi gives at most 3 courses at A01. The second tuple says that if the room which Ishihara uses is either at room B02 or C03, then Ishihara has two courses, and otherwise, Ishihara has no courses.

Next, we also provide two semantics [16] to these models. One semantics is closed world assumption (called CWA), which means "invisible tuples do not exist." For example, Figure 3 under CWA describes that no room other than A01, B02, and C03 is used. The other is open world assumption (called OWA), which means that "the existence of invisible tuples is unknown." For example, Figure 3 under OWA describes that some room other than A01, B02, and C03 may be used.

Then, we consider operations which are often used in a query to a database. In this paper, the closure properties of algebraic operations, selection σ, projection π, union $\cup$, product $\times$, difference $-$ and unique μ, and their inverses on each model under these semantics are investigated.

The closure properties of forward operations are important since the answer of a query to incomplete information should be represented under the same data model. The closure properties of inverse operations are also important when the incomplete information on a database is derived from a query to the database and its answer. For example, suppose that the user has issued a query q to the database D and obtained the answer A. Then, Disclosure Monitor mentioned above computes the user's knowledge (a) and (b), i.e., the incomplete information on D determined by q and A. Roughly speaking, this computation is accomplished by applying the inverse of q to A. Therefore, in order for the incomplete information on D to be represented under a data model, the inverse of every operation in q is desirable to be closed on the same data model.

Our results are summarized in Table 1. The results newly obtained in this paper are represented by bold letters. The inverse of product $\times$ under CWA, the inverse of selection σ under OWA and the inverse of unique μ under CWA are not closed on V-bags in Table 1 (a) and (b), whereas they are closed on GV-bags in Table 1 (c) and (d).

1.3 Related works

There are some related works on bag-based data models for incomplete information. In [14], a partial order representing a degree of incompleteness between bags is introduced.

In the model of [14], partial information such as "A's value is unknown" can be captured, but ambiguous information such as "A's value is B or C" cannot be represented naturally. On the other hand, in our model, ambiguous information can be naturally represented. In [12], aggregate queries on C-tables are defined. Then, it is shown that the aggregate queries on C-tables are closed. In that model, aggregate values are represented by the values of a special attribute. On the other hand, our model is purely a bag-based model.

In [15], a standard bag language, called BQL (additive union, monus, max, min, eq, member and unique), is proposed, where additive union, monus and unique are the same as union, difference and unique in this paper. Monus can express all operations in BQL other than additive union and unique, which are independent of the rest of other operations. Therefore, in this paper, the closure properties of these three important operations are investigated. Moreover, in [15], an element of a bag is not a tuple but an atomic fact while, in this paper, an element of a bag is a tuple. Therefore, investigating the closure properties of selection, projection, and product is also important.

1.4 Organization

The rest of this paper is organized as follows. In Section 2, we define bag-based databases, V-bags, GV-bags, CWA and OWA . In Section 3, we prove the closure properties of unique and its inverse on V-bags under CWA. In Section 4, we prove the closure properties on GV-bags under both CWA and OWA. Lastly, in Section 5, we provide the summary and future work.

2. Definitions

2.1 Bag-based databases

In this section, we extend the definition of ordinary relational databases in [2] to bag-based databases. Although *bag*-based databases are not *relational*, we borrow the terminology of relational databases.

Definition 1 A *relation schema* R is a set of attributes. We assume that the domain of every attribute in R is the set N of non-negative integers (or every attribute value is encoded by a non-negative integer). A *tuple* t over R is a function from R to N. Let $t(A)$ denote the value of $A \in R$ in t. A *relational instance* D over R is a function from the set of tuples over R to N such that $\{t \mid D(t) \neq 0\}$ is a finite set. A *database schema* $\mathbf{R}$ is a finite sequence $\langle R_1, \ldots, R_n \rangle$ of relation schemas. A *database instance* over a database schema $\mathbf{R} = \langle R_1, \ldots, R_n \rangle$ is a sequence $\langle D_1, \ldots, D_n \rangle$, where each D_i is a relational instance of R_i.

For a tuple t over R and $X \subseteq R$, let $t[X]$ denote the function obtained by restricting the domain of t to X. Let $dom(D) = \{t \mid D(t) \neq 0\}$. If $D_1(t) \leqq D_2(t)$ for an arbitrary tuple t, then we write $D_1 \subseteq D_2$.

Birthday	Name	
June 10	Sato	$\mapsto 3$
May 5	Tanaka	$\mapsto 2$

Figure 4 A relational instance D in Example 3.

Example 3 Suppose that there are three people whose surnames are Sato and the dates of the birth are June 10. Moreover, assume that there are two other people whose surnames are Tanaka and the dates of the birth are May 5. Then, the relational instance D representing these facts is as follows (see Fig. 4):

$$D(t) = \begin{cases} 3 & \text{if } t = (\text{June } 10, \text{Sato}), \\ 2 & \text{if } t = (\text{May } 5, \text{Tanaka}), \\ 0 & \text{otherwise.} \end{cases}$$

In Fig. 4, the numbers of tuples are shown to the right of $\mapsto$. By definition, $dom(D) = \{(\text{June } 10, \text{Sato}), (\text{May } 5, \text{Tanaka})\}$.

Definition 2 We define *selection*, *projection*, *union*, *product*, *difference* and *unique* as follows:
Selection $\sigma_C(D)$: Let D be a relational instance over a relation schema R and C be a selection condition. For every tuple t over R, we define $\sigma_C(D)(t)$ as follows:

$$\sigma_C(D)(t) = \begin{cases} D(t) & \text{if } t \text{ satisfies } C, \\ 0 & \text{otherwise.} \end{cases}$$

Projection $\pi_X(D)$: Let D be a relational instance over a relation schema R and let $X \subseteq R$. For every tuple t over X, we define $\pi_X(D)(t)$ as follows:

$$\pi_X(D)(t) = \sum_{t' : t'[X] = t} D(t').$$

Union $D_1 \cup D_2$: Let D_1 and D_2 be relational instances over a relation schema R. For every tuple t over R, we define $(D_1 \cup D_2)(t)$ as follows:

$$(D_1 \cup D_2)(t) = D_1(t) + D_2(t).$$

Product $D_1 \times D_2$: Let D_1 and D_2 be relational instances over relation schemas R_1 and R_2, respectively, such that $R_1 \cap R_2 = \emptyset$. For every t_1 over R_1 and every t_2 over R_2, we define $(D_1 \times D_2)(t_1 t_2)$ as follows:

$$(D_1 \times D_2)(t_1 t_2) = D_1(t_1) \times D_2(t_2),$$

where $t_1 t_2$ denotes the tuple over $R_1 \cup R_2$ such that $t_1 t_2[R_1] = t_1$ and $t_1 t_2[R_2] = t_2$, and $\times$ in the right-hand side denotes the arithmetic multiplication.

Difference $D_1 - D_2$: Let D_1 and D_2 be relational instances over a relation schema R. For every tuple t, we define $(D_1 - D_2)(t)$ as follows:

$$(D_1 - D_2)(t) = \max(D_1(t) - D_2(t), 0),$$

where $-$ in the right-hand side denotes the arithmetic subtraction.

Table 2 Conditional expressions defined as macros on non-negative integer expressions.

conditional expressions	non-negative integer expressions
true	1
false	0
$a = b$	$1 \dot{-} ((a \dot{-} b) + (b \dot{-} a))$
$a \leqq b$	$1 \dot{-} (a \dot{-} b)$
$\bigvee_i c_i$	$1 \dot{-} \left(1 \dot{-} \sum_i c_i\right)$
$\bigwedge_i c_i$	$\prod_i c_i$
$\neg c_1$	$1 \dot{-} c_1$
if c_1 then a else b	$c_1 \times a + (1 \dot{-} c_1) \times b$

Unique $\mu(D)$: Let D be a relational instance over a relation schema R. For every tuple t, we define $\mu(D)(t)$ as follows:

$$\mu(D)(t) = \begin{cases} 1 & \text{if } D(t) \geqq 1, \\ 0 & \text{otherwise.} \end{cases}$$

2. 2 V-bags

Let V be a set of *variables*. Let $\dot{-}$ denote the difference operation on non-negative integers, i.e.,

$$a \dot{-} b = \begin{cases} a - b & \text{if } a \geqq b, \\ 0 & \text{otherwise.} \end{cases}$$

A *non-negative integer expression* is an expression consisting of non-negative integers, variables, and operators $+$, $\times$ and $\dot{-}$. Hereafter, we use $\sum_{i=1}^{n} a_i$ and $\prod_{i=1}^{n} a_i$ as shorthand notations for $a_1 + \cdots + a_n$ and $a_1 \times \cdots \times a_n$, respectively.

A *V-tuple* u over R is a total function from R to $N \cup V$. A *V-bag* E over R is a total function from the set of V-tuples over R to the set of non-negative integer expressions such that $\{u \mid E(u) \not\equiv 0\}$ is finite, where $E(u) \not\equiv 0$ means that the expression of $E(u)$ is not literally 0. Let $dom(E) = \{u \mid E(u) \not\equiv 0\}$. For example, let $E(u) = 0$ and $E(u') = x \dot{-} x$. Then, $u \notin dom(E)$ but $u' \in dom(E)$.

In Table 2, we introduce *conditional expressions* defined as macros on non-negative integer expressions, where a and b are arbitrary non-negative integer expressions and c_i is an arbitrary non-negative integer expression such that the value of c_i is equal to 0 or 1.

A *valuation* is defined as a function from V to N. The domain of a valuation ν is naturally extended as follows:

- For each constant $a \in N$, let $\nu(a) = a$.
- For non-negative integer expressions $x + y$, $x \times y$ and $x \dot{-} y$, let $\nu(x+y) = \nu(x) + \nu(y)$, $\nu(x \times y) = \nu(x) \times \nu(y)$ and $\nu(x \dot{-} y) = \nu(x) \dot{-} \nu(y)$.
- For a V-tuple u over X, let $\nu(u)$ be a tuple over X satisfying that for each $A \in X$, $(\nu(u))(A) = \nu(u(A))$.
- For a V-bag E and a tuple t,

$$\nu(E)(t) = \sum_{u:\nu(u)=t} \nu(E(u)).$$

We provide two semantics. The first semantics is *Closed World Assumption* (CWA). The other is *Open World Assumption* (OWA).

Definition 3 The set $rep_C(\langle E_1, \ldots, E_n\rangle)$ of database instances represented by $\langle E_1, \ldots, E_n\rangle$ under *CWA* is defined as follows:

$$rep_C(\langle E_1, \ldots, E_n\rangle) = \{\langle D_1, \ldots, D_n\rangle \mid D_1 = \nu(E_1), \ldots, D_n = \nu(E_n) \text{ for some valuation } \nu\}.$$

The set $rep_O(\langle E_1, \ldots, E_n\rangle)$ of database instances represented by $\langle E_1, \ldots, E_n\rangle$ under *OWA* is defined as follows:

$$rep_O(\langle E_1, \ldots, E_n\rangle) = \{\langle D_1, \ldots, D_n\rangle \mid D_1 \supseteqq \nu(E_1), \ldots, D_n \supseteqq \nu(E_n) \text{ for some valuation } \nu\}.$$

If $n = 1$, then we write $rep_C(E_1)$ and $rep_O(E_1)$ instead of $rep_C(\langle E_1\rangle)$ and $rep_O(\langle E_1\rangle)$, respectively.

Definition 4 Let q be an operation on bag-based databases with n inputs and m outputs. The operation q is *closed* on V-bags under CWA if for any sequence $\langle E_1, \ldots, E_n\rangle$ of V-bags, there is a sequence $\langle E'_1, \ldots, E'_m\rangle$ of V-bags such that

$$rep_C(\langle E'_1, \ldots, E'_m\rangle) = \{q(\langle D_1, \ldots, D_n\rangle) \mid \langle D_1, \ldots, D_n\rangle \in rep_C(\langle E_1, \ldots, E_n\rangle)\}.$$

The inverse of operation q is *closed* on V-bags under CWA if for any $\langle E'_1, \ldots, E'_m\rangle$, there is $\langle E_1, \ldots, E_n\rangle$ of V-bags such that

$$rep_C(\langle E_1, \ldots, E_n\rangle) = \{\langle D_1, \ldots, D_n\rangle \mid q(\langle D_1, \ldots, D_n\rangle) \in rep_C(\langle E'_1, \ldots, E'_m\rangle)\}.$$

Closure properties on V-bags under OWA is defined in the same way.

2. 3 GV-bags

We propose a supermodel of V-bags called *GV-bags*. A GV-bag is a pair $(E, \mathcal{G})$ of a V-bag E and a global condition $\mathcal{G}$. A sequence of GV-bags with a same global condition $\mathcal{G}$ $\langle(E_1, \mathcal{G}), (E_2, \mathcal{G}), \ldots, (E_n, \mathcal{G})\rangle$ is rewritten as $(\langle E_1, \ldots, E_n\rangle, \mathcal{G})$. If $n = 1$, then we write $(E_1, \mathcal{G})$ instead of $(\langle E_1\rangle, \mathcal{G})$. A global condition is a Boolean expression consisting of conditional expressions shown in the left column of Table 2.

Definition 5 The set $rep_C^{\mathcal{G}}(\langle E_1, \ldots, E_n\rangle, \mathcal{G})$ of database instances represented by $(\langle E_1, \ldots, E_n\rangle, \mathcal{G})$ under CWA is defined as follows:

$$rep_C^{\mathcal{G}}(\langle E_1, \ldots, E_n\rangle, \mathcal{G}) = \{\langle D_1, \ldots, D_n\rangle \mid D_1 = \nu(E_1), \ldots, D_n = \nu(E_n) \text{ for some valuation } \nu \text{ such that } \nu(\mathcal{G}) = \text{true}\}.$$

The set $rep_O^{\mathcal{G}}(\langle E_1, \ldots, E_n\rangle, \mathcal{G})$ of database instances represented by $(\langle E_1, \ldots, E_n\rangle, \mathcal{G})$ under OWA is defined as follows:

$$rep_O^{\mathcal{G}}(\langle E_1, \ldots, E_n\rangle, \mathcal{G}) = \{\langle D_1, \ldots, D_n\rangle \mid$$

$D_1 \supseteqq \nu(E_1), \ldots, D_n \supseteqq \nu(E_n)$

for some valuation ν such that $\nu(\mathcal{G}) = \text{true}\}$.

If $n = 1$, then we write $rep^G_C(E_1, \mathcal{G})$ and $rep^G_O(E_1, \mathcal{G})$ instead of $rep^G_C(\langle E_1 \rangle, \mathcal{G})$ and $rep^G_O(\langle E_1 \rangle, \mathcal{G})$, respectively.

Note that if $\nu(\mathcal{G}) = \text{false}$ for any valuation ν, then $rep^G(E, \mathcal{G}) = \emptyset$, where $\emptyset$ means there is no instance in $rep^G(E, \mathcal{G})$. On the other hand, if $\nu(E(u)) = 0$ for any $u \in dom(E)$ and any valuation ν such that $\nu(\mathcal{G}) = \text{true}$, then $rep^G(E, \mathcal{G}) = \{\emptyset\}$, where $\{\emptyset\}$ means that $rep^G(E, \mathcal{G})$ contains only the empty instance.

The closure properties of six operations on GV-bags are defined in the same way as V-bags.

3. Closure properties of V-bags

In this section, we show that unique under CWA is closed (Theorem 1). Then, we show that the inverse of unique under CWA is not closed (Theorem 2). The proofs of the case of OWA are similar to the case of CWA.

Unique on V-bags under CWA is defined as follows.

Definition 6 Let E be a V-bag over a relation schema R. Fix some total order $\preceq$ on $dom(E)$. For each V-tuple u, we define $\mu(E)(u)$ as follows:

$$\mu(E)(u) = \begin{cases} \displaystyle\sum_{F \subseteqq dom(E)} \text{“if } \Phi^u_F \text{ then } \Xi^u_F \text{ else 0”} \\ \qquad \text{if } u \in dom(E), \\ 0 \qquad \text{otherwise,} \end{cases}$$

where

$$\Phi^u_F = \left(\bigwedge_{u' \in F} \bigwedge_{A \in R} (u[A] = u'[A]) \right) \wedge \left(\bigwedge_{u' \in dom(E) - F} \bigvee_{A \in R} \neg(u[A] = u'[A]) \right),$$

and

$$\Xi^u_F = \begin{cases} 1 \dot{-} \left(1 \dot{-} \displaystyle\sum_{u' \in F} E(u') \right) \\ \qquad \text{if } u \text{ is the largest in } F \text{ w.r.t. } \preceq, \\ 0 \qquad \text{otherwise.} \end{cases}$$

Theorem 1 Unique is closed on V-bags under CWA, that is, $rep_C(\mu(E)) = \{\mu(D) \mid D \in rep_C(E)\}$.

Proof: Let E be a V-bag over a relation schema R. We prove that

$$\forall \nu \forall u, \nu(\mu(E))(\nu(u)) = \mu(\nu(E))(\nu(u)).$$

The theorem is immediately derived from this equation and the definition of CWA. We will prove the equation by showing that both-hand side can be rewritten as $1 \dot{-} (1 \dot{-} \nu(E)(\nu(u)))$. First, we consider the Left-hand side. From the construction of Φ^u_F, it can be concluded that for each pair of ν and u, there is exactly one $F^{\nu,u}$ such that $\nu(\Phi^u_{F^{\nu,u}}) = \text{true}$, namely,

$$F^{\nu,u} = \{u' \in dom(E) \mid \nu(u') = \nu(u)\}.$$

Therefore, we have

$$\begin{aligned} &\nu(\mu(E))(\nu(u)) \\ &= \sum_{u' : \nu(u') = \nu(u)} \nu(\mu(E(u'))) \\ &= \sum_{u' : \nu(u') = \nu(u)} \nu \left(\sum_{F \subseteqq dom(E)} \text{“if } \Phi^{u'}_F \text{ then } \Xi^{u'}_F \text{ else 0”} \right) \\ &= \sum_{u' : \nu(u') = \nu(u)} \sum_{F \subseteqq dom(E)} \text{“if } \nu\left(\Phi^{u'}_F\right) \text{ then } \nu\left(\Xi^{u'}_F\right) \text{ else 0”} \\ &= \sum_{u' : \nu(u') = \nu(u)} \nu\left(\Xi^{u'}_{F^{\nu,u'}}\right) \\ &= \nu \left(1 \dot{-} \left(1 \dot{-} \sum_{u \in F^{\nu,u}} E(u) \right) \right), \end{aligned}$$

since u' varies over $F^{\nu,u'}$ and u is in $F^{\nu,u'}$. Therefore, we have

$$\begin{aligned} \nu(\mu(E))(\nu(u)) &= \nu \left(1 \dot{-} \left(1 \dot{-} \sum_{u \in F^{\nu,u}} E(u) \right) \right) \\ &= 1 \dot{-} (1 \dot{-} \nu(E)(\nu(u))). \end{aligned}$$

This equation indicates that if $\nu(E)(\nu(u)) \geqq 1$, the number of $\nu(u)$ is 1, and otherwise, the number of $\nu(u)$ is 0. Now, we consider the Right-hand side. From Definition 2 and the property of $\dot{-}$, we obtain

$$\begin{aligned} \mu(\nu(E))(\nu(u)) &= \begin{cases} 1 & \text{if } \nu(E)(\nu(u)) \geqq 1, \\ 0 & \text{otherwise,} \end{cases} \\ &= 1 \dot{-} (1 \dot{-} \nu(E)(\nu(u))). \end{aligned}$$

□

Now, we show that the inverse of unique on V-bags under CWA is not closed.

Theorem 2 The inverse of unique is not closed on V-bags under CWA.

Proof: We assume that the inverse of unique is closed on V-bags under CWA, and derive a contradiction. Let a be a non-negative constant. Let E' be a V-bag over a relation schema $\{A\}$ with $dom(E') = \{u\}$, $u(A) = a$ and $E'(u) = 2$ (see Figure 5). From the assumption that the inverse of unique is closed, there is a V-bag E such that

$$rep_C(E) = \{D \mid \mu(D) \in rep_C(E')\}.$$

For $D' \in rep_C(E')$, there is an only tuple t such that

A	
a	$\mapsto 2$

Figure 5 E' in Theorem 2.

$D'(t) = 2$, and hence, there is no D such that $\mu(D) = D'$. Thus,

$$rep_C(E) = \{D \mid \mu(D) \in rep_C(E')\} = \emptyset.$$

On the other hand, by the definition of CWA,

$$rep_C(E) = \{D \mid D = \nu(E) \text{ for some valuation } \nu\}.$$

cannot be empty. Hence, the inverse of unique is not closed under CWA. $\square$

4. Closure properties of GV-bags

In this section, we show that the proofs of the inverse of selection under OWA, the inverse of product under CWA and the inverse of unique under CWA are closed (Theorems 3, 4 and 5). The proof of the inverse of unique under OWA is similar to the CWA case. The proofs of the other operations are similar to the case of V-bags. They are omitted because of the space limitation.

Now, we prove that inverse of selection under OWA is closed.

Definition 7 Let $(E, \mathcal{G})$ be a GV-bag over a relation schema R and C be a selection condition. Let C_u be a condition such that every attribute A in C is replaced with $u(A)$ (e.g., if C is "$A = a$" then C_u is "$u(A) = a$"). Let $(E', \mathcal{G}') = \sigma_C^{-1}(E, \mathcal{G})$. We define E' as E and $\mathcal{G}'$ as follows:

$$\mathcal{G}' = \left(\bigwedge_{u \in dom(E)} C_u \right) \wedge \mathcal{G}.$$

Theorem 3 The inverse of selection is closed under OWA, that is, $\sigma_C(D) \in rep_O^G(E, \mathcal{G})$ if and only if $D \in rep_O^G(\sigma_C^{-1}(E, \mathcal{G}))$.

Proof: *If part.* Let $(E, \mathcal{G})$ and C be a GV-bag over a relation schema R and a selection condition, respectively. Let $(E', \mathcal{G}') = \sigma_C^{-1}(E, \mathcal{G})$. Consider an instance D such that $D \in rep_O^G(E', \mathcal{G}')$. There must be a valuation ν' such that $D \supseteqq \nu'(E')$ and $\nu'(\mathcal{G}') = $ true. Hence, for each tuple $u' \in dom(E')$, $\nu'(u')$ satisfies C_u, since $\nu'(\mathcal{G}') = $ true. Therefore, we have $\sigma_C(D) \supseteqq \nu'(E')$. We obtain $\sigma_C(D) \supseteqq \nu'(E)$, because $E' = E$. Hence, we have $\sigma_C(D) \supseteqq \nu'(E) \in rep_O^G(E, \mathcal{G})$.

Only if part. Let $(E, \mathcal{G})$ and C be a GV-bag over a relation schema R and a selection condition, respectively. Let D be an arbitrary instance such that $\sigma_C(D) \in rep_O^G(E, \mathcal{G})$. Then, there is a valuation ν such that $\sigma_C(D) \supseteqq \nu(E)$. Therefore, all tuples in $\nu(E)$ satisfy C. Let $(E', \mathcal{G}') = \sigma_C^{-1}(E, \mathcal{G})$. We obtain $\nu(\mathcal{G}')$ is true, and $\nu(E') = \nu(E) \subseteqq \sigma_C(D) \subseteqq D$ because $E = E'$. Hence, we have $D \in rep_O^G(E', \mathcal{G}') = rep_O^G(\sigma_C^{-1}(E, \mathcal{G}))$. $\square$

Now, the inverse of product on GV-bags under CWA is defined as follows.

Definition 8 Let X and Y be relation schemas such that $X \cap Y = \emptyset$. Let $(E, \mathcal{G})$ be a GV-bag over the relation schema $X \cup Y$. Let $G_1 = \{u[X] \mid u \in dom(E)\}$ and $G_2 = \{u[Y] \mid u \in dom(E)\}$. We define $\times^{-1}(E, \mathcal{G})$ as the following pair of GV-bags $(E_1, \mathcal{G}')$ over X and $(E_2, \mathcal{G}')$ over Y. First, the global condition $\mathcal{G}'$ is defined as follows:

$$\mathcal{G}' = \bigwedge_{u_1 u_2 \in dom(E)} \left(\bigwedge_{F_1 \subseteqq G_1} \bigwedge_{F_2 \subseteqq G_2} \Omega_{F_1 F_2} \right) \wedge \mathcal{G},$$

where

$$\Omega_{F_1 F_2} = \text{"if } \Phi_{F_1 F_2} \text{ then } \sum_{u_1 \in F_1, u_2 \in F_2} (E, \mathcal{G})(u_1 u_2) = S_{F_1 F_2}\text{,"}$$

$$\begin{aligned}\Phi_{F_1 F_2} = {} & \left(\bigwedge_{u' \in F_1} \bigwedge_{A \in X} (u_1(A) = u'(A)) \right) \\ & \wedge \left(\bigwedge_{u' \in G_1 - F_1} \bigvee_{A \in X} \neg(u_1(A) = u'(A)) \right) \\ & \wedge \left(\bigwedge_{u' \in F_2} \bigwedge_{A \in Y} (u'(A) = u_2(A)) \right) \\ & \wedge \left(\bigwedge_{u' \in G_2 - F_2} \bigvee_{A \in Y} \neg(u'(A) = u_2(A)) \right),\end{aligned}$$

$$S_{F_1 F_2} = \sum_{u_1 \in F_1} (E_1, \mathcal{G}')(u_1) \times \sum_{u_2 \in F_2} (E_2, \mathcal{G}')(u_2).$$

For each V-tuple u, introduce new variables x_{u_X} and y_{u_Y} not appearing in $dom(E)$, and define $(E_1, \mathcal{G}')$ and $(E_2, \mathcal{G}')$ as follows:

$$(E_1, \mathcal{G}')(u_X) = \begin{cases} x_{u_X} & \text{if } u_X \in G_1, \\ 0 & \text{otherwise,} \end{cases}$$

$$(E_2, \mathcal{G}')(u_Y) = \begin{cases} y_{u_Y} & \text{if } u_Y \in G_2, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 4 The inverse of product is closed on GV-bags under CWA, that is, $D_1 \times D_2 \in rep_C^G(E, \mathcal{G})$ if and only if $\langle D_1, D_2 \rangle \in rep_C^G(\times^{-1}(E, \mathcal{G}))$.

Proof: *If part.* Let $(E, \mathcal{G})$ be a GV-bag over a relation schema $X \cup Y$. Let $(\langle E_1, E_2 \rangle, \mathcal{G}')$ be $\times^{-1}(E, \mathcal{G})$. Consider instances D_1 and D_2 such that $\langle D_1, D_2 \rangle \in rep_C^G(\langle E_1, E_2 \rangle, \mathcal{G}')$. There must be a valuation ν' such that $\nu'(E_1) = D_1$, $\nu'(E_2) = D_2$ and $\nu'(\mathcal{G}') = $ true. For ν', $t_1 \in D_1$, and $t_2 \in D_2$, define

$$F_1^{\nu',t_1} = \{u_1 \in G_1 \mid \nu'(u_1) = t_1\},$$
$$F_2^{\nu',t_2} = \{u_2 \in G_2 \mid \nu'(u_2) = t_2\}.$$

Then,

$$\begin{aligned}\nu'(E_1)(t_1) &= \nu'\left(\sum_{u_1:\nu'(u_1)=t_1} (E_1,\mathcal{G})(u_1)\right)\\ &= \sum_{u_1:\nu'(u_1)=t_1} \nu'((E_1,\mathcal{G})(u_1))\\ &= \sum_{u_1\in F_1^{\nu',t_1}} \nu'(x_{u_1}).\end{aligned}$$

Similarly, $\nu'(E_2)(t_2) = \sum_{u_2\in F_2^{\nu',t_2}} \nu'(y_{u_2})$.

Since $\nu'(\mathcal{G}') = \text{true}$, we obtain

$$\begin{aligned}&\sum_{u_1\in F_1^{\nu',t_1}} \nu'(x_{u_1}) \times \sum_{u_2\in F_2^{\nu',t_2}} \nu'(y_{u_2})\\ &\quad= \sum_{u_1\in F_1^{\nu',t_1}, u_2\in F_2^{\nu',t_2}} \nu'((E,\mathcal{G})(u_1u_2))\\ &\quad= \sum_{u_1u_2:\nu'(u_1u_2)=t_1t_2} \nu'((E,\mathcal{G})(u_1u_2))\\ &\quad= \nu'(E)(t_1t_2).\end{aligned}$$

Therefore, we have $\nu'(E_1)(t_1) \times \nu'(E_2)(t_2) = \nu'(E)(t_1t_2)$. Thus, if $D_1 = \nu'(E_1)$ and $D_2 = \nu'(E_2)$, then $D_1 \times D_2 = \nu'(E)$. By the definition of CWA, $D_1 \times D_2 \in rep_C^G(E,\mathcal{G})$.

Only if part. Let $(E,\mathcal{G})$ be a GV-bag over a relation schema $X \cup Y$, and D_1 and D_2 be arbitrary instances such that $D_1 \times D_2 \in rep_C^G(E,\mathcal{G})$. Then, there is a valuation ν such that $\nu(E) = D_1 \times D_2$, and $\nu(\mathcal{G}) = \text{true}$. Therefore, for each t_1t_2 in $\nu(E)$, $\nu(E)(t_1t_2) = D_1(t_1) \times D_2(t_2)$. Let $(\langle E_1, E_2\rangle, \mathcal{G}')$ be $\times^{-1}(E,\mathcal{G})$. In what follows, we construct a valuation ν' such that for each pair of tuples $t_1 \in dom(D_1)$ and $t_2 \in dom(D_2)$,

$$\nu'(E_1)(t_1) = D_1(t_1),$$
$$\nu'(E_2)(t_2) = D_2(t_2).$$

Such ν' satisfies $D_1 = \nu'(E_1)$, $D_2 = \nu'(E_2)$ and $\nu'(\mathcal{G}') = \text{true}$.

ν' is defined on the variables appearing in $dom(E_1)$ or $dom(E_2)$ and the new variables x_{u_X} and y_{u_Y}. For each variable z appearing in $dom(E_1)$ or $dom(E_2)$, let $\nu'(z) = \nu(z)$. Define

$$F_1^{\nu',t_1} = \{u_1 \in G_1 \mid \nu'(u_1) = t_1\},$$
$$F_2^{\nu',t_2} = \{u_2 \in G_2 \mid \nu'(u_2) = t_2\}.$$

We prove the existence of ν' such that

$$\sum_{u_1\in F_1^{\nu',t_1}} \nu'(x_{u_1}) = D_1(t_1),$$
$$\sum_{u_2\in F_2^{\nu',t_2}} \nu'(y_{u_2}) = D_2(t_2).$$

From the construction of F_1^{ν',t_1} and F_2^{ν',t_2}, if $t_1 \neq t_1'$, then $F_1^{\nu',t_1} \cap F_1^{\nu',t_1'} = \emptyset$. Therefore, each variable in G_1 and G_2 appears only once. Hence, we can easily construct the valuation ν' satisfying above equations. Then,

$$\begin{aligned}&\sum_{u_1\in F_1^{\nu',t_1}, u_2\in F_2^{\nu',t_2}} \nu'((E,\mathcal{G})(u_1u_2))\\ &= \sum_{u_1u_2:\nu'(u_1u_2)=t_1t_2} \nu'((E,\mathcal{G})(u_1u_2)) = \nu'(E)(t_1t_2).\end{aligned}$$

Hence, we have $\langle D_1, D_2\rangle \in rep_C^G(\langle E_1, E_2\rangle, \mathcal{G}) = rep_C^G(\times^{-1}(E,\mathcal{G}))$. □

Now, we prove that inverse of unique under CWA is closed.

Definition 9 Let $(E,\mathcal{G})$ be a GV-bag over a relation schema R. Let $(E',\mathcal{G}') = \mu^{-1}(E,\mathcal{G})$. First, the global condition $\mathcal{G}'$ is defined as follows:

$$\mathcal{G}' = \bigwedge_{u\in dom(E)} \left(\bigwedge_{F\subseteqq dom(E)} \Omega_F\right) \wedge \mathcal{G},$$

where

Ω_F = "if Φ_F then $\sum_{u\in F}(E,\mathcal{G})(u) = 1$,"

$$\begin{aligned}\Phi_F = &\left(\bigwedge_{u'\in F}\bigwedge_{A\in X}(u[A] = u'[A])\right)\\ &\wedge\left(\bigwedge_{u'\in dom(E)-F}\bigvee_{A\in X}\neg(u[A] = u'[A])\right).\end{aligned}$$

For each V-tuple u, introduce a new variable x_u not appearing in $dom(E)$, and define $(E',\mathcal{G}')$ as follows:

$$(E',\mathcal{G}')(u) = \begin{cases}\text{"if } (E,\mathcal{G})(u) \text{ then } x_u + 1 \text{ else } 0\text{"} & \text{if } u \in dom(E),\\ 0 & \text{otherwise.}\end{cases}$$

Theorem 5 The inverse of unique is closed under CWA, that is, $\mu(D) \in rep_C^G(E,\mathcal{G})$ if and only if $D \in rep_C^G(\mu^{-1}(E,\mathcal{G}))$.

Proof: *If part.* Let $(E,\mathcal{G})$ be a GV-bag over a relation schema R. Let $(E',\mathcal{G}') = \mu^{-1}(E,\mathcal{G})$. Consider an instance D such that $D \in rep_C^G(E',\mathcal{G}')$. There must be a valuation ν' such that $D = \nu'(E')$ and $\nu'(\mathcal{G}') = \text{true}$. For ν' and $t \in dom(D)$, define

$$F^{\nu',t} = \{u \in dom(E) \mid \nu'(u) = t\}.$$

Then,

$$\begin{aligned}\mu(D)(t) &= \mu(\nu'(E'))(t)\\ &= 1 \dot{-} \left(1 \dot{-} \nu'\left(\sum_{u:\nu'(u)=t}(E',\mathcal{G}')(u)\right)\right)\end{aligned}$$

$$= 1 \dot{-} \left(1 \dot{-} \nu' \left(\sum_{u \in F^{\nu',t}} S\right)\right),$$

where

$$S = \text{“if } (E, \mathcal{G})(u) \text{ then } x_u + 1 \text{ else } 0\text{”}.$$

Since $\nu'(\mathcal{G}') = \text{true}$,

$$\nu' \left(\sum_{u \in F^{\nu',t}} (E, \mathcal{G})(u)\right) = 1.$$

Hence, for one $u' \in F^{\nu',t}$, we have $\nu'((E, \mathcal{G})(u')) = 1$, and for the others $u'' \in F^{\nu',t}$, we have $\nu'((E, \mathcal{G})(u'')) = 0$. Thus, we obtain $\mu(D)(t) = 1 \dot{-} (1 \dot{-} (\nu'(x_{u'} + 1))) = 1$. Therefore, we have $\mu(D) = \nu'(E)$. Hence, we have $\mu(D) = \nu'(E) \in rep_C^G(E, \mathcal{G})$.

Only if part. Let $(E, \mathcal{G})$ be a GV-bag over a relation schema R. Let D be an arbitrary instance such that $\mu(D) \in rep_C^G(E, \mathcal{G})$. Then, there is a valuation ν such that $\mu(D) = \nu(E)$ and $\nu(\mathcal{G}) = \text{true}$. We obtain $\nu(\mathcal{G}') = \text{true}$ because, for each $\nu(u)$ in $\nu(E)$, $\nu(E)(\nu(u))(= \mu(D)(\nu(u))) = 1$. Let $(E', \mathcal{G}') = \mu^{-1}(E, \mathcal{G})$. In what follows, we construct a valuation ν' such that for each tuple $t \in dom(D)$,

$$\nu'(E')(t) = D(t).$$

ν' is defined on the variables appearing in $dom(E')$ and the new variables x_u. For each variable z appearing in $dom(E')$, let $\nu'(z) = \nu(z)$. Such ν' satisfies $\nu'(\mathcal{G}') = \text{true}$.

Define

$$F^{\nu',t} = \{u \in dom(E) \mid \nu'(u) = t\}.$$

Then, we must prove the existence of ν' such that

$$\nu'(E')(t) = \sum_{u \in F^{\nu',t}} \nu'(\text{“if } (E, \mathcal{G})(u) \text{ then } x_u + 1 \text{ else } 0\text{”}) = D(t).$$

Since $\nu'(\mathcal{G}') = \text{true}$, for one $u' \in F^{\nu',t}$, we have $\nu'((E, \mathcal{G})(u')) = 1$, and for the others $u'' \in F^{\nu',t}$, we have $\nu'((E, \mathcal{G})(u'')) = 0$. From the construction of $F^{\nu',t}$, if $t \neq t'$, then $F^{\nu',t} \cap F^{\nu',t'} = \emptyset$. From Definition 9, for each $u \in dom(E)$, x_u in $dom(E')$ appears only once. Hence, we can easily construct the valuation ν' satisfying above equations. Hence, we have $D \in rep_C^G(E', \mathcal{G}') = rep_C^G(\mu^{-1}(E, \mathcal{G}))$. □

5. Conclusion and future work

We have proposed a bag-based model, called GV-bags, for representing incomplete information. Then, we have proved the closure properties of unique and its inverse on V-bags under both CWA and OWA, and the closure properties of six operations and their inverses on GV-bags under both CWA and OWA.

As a future work, we will evaluate the computational complexity of the operations and the sizes of V-bags and GV-bags after the operations. Also, we will investigate a new submodel of GV-bags where the usage of variables for representing the number of tuples is somewhat restricted. In this model, we expect that inverse of projection is closed, which is not closed on both V-bags and GV-bags under both CWA and OWA.

References

[1] S. Abiteboul, and O. M. Duschka, “Complexity of Answering Queries Using Materialized Views,” *Proc. 17th ACM PODS*, pp. 254–263, 1998.

[2] S. Abiteboul, R. Hull, and V. Vianu, *Foundations of Databases*, Addison-Wesley, 1995.

[3] S. Abiteboul, P. Kanellakis, and G. Grahne, “On the Representation and Querying of Sets of Possible Worlds,” *Proc. ACM SIGMOD*, pp. 34–48, 1987.

[4] S. Abiteboul, and G. Grahne, “Update Semantics for Incomplete Databases,” *Proc. 11th Internat. Conf. on VLDB*, pp. 1–12, 1985.

[5] A. Brodsky, C. Farkas, and S. Jajodia, “Secure Databases: Constraints, Inference Channels, and Monitoring Disclosures,” *IEEE TKDE*, 12, 6, pp. 900–919, 2000.

[6] D. Calvanese, G. De Giacomo, M. Lenzerini, and M. Y. Vardi, “Answering Regular Path Queries Using Views,” *Proc. 16th ICDE*, pp. 389–398, 2000.

[7] E. F. Codd, “Understanding Relations,” *FDT Bulletin of ACM SIGMOD*, 7, 3/4, pp. 23–28, 1975.

[8] G. Grahne, “Dependency Satisfaction in Databases with Incomplete Information,” *Proc. 10th Internat. Conf. on VLDB*, pp. 37–45, 1984.

[9] G. Grahne, *The Problem of Incomplete Information in Relational Databases*, LNCS 554, Springer-Verlag, 1991.

[10] A. Y. Halevy, “Answering Queries Using Views: A Survey,” *VLDB Journal*, 10, pp. 270–294, 2001.

[11] T. Imielinski, and W. Lipski, “Incomplete Information in Relational Databases,” *J. ACM*, 31, 4, pp. 761–791, 1984.

[12] J. Lechtenbörger, H. Shu, and G. Vossen, “Aggregate Queries over Conditional Tables,” *J. Intelligent Information Systems*, 19, 3, pp. 343–362, 2002.

[13] A. Y. Levy, A. O. Mendelzon, Y. Sagiv, and D. Srivastava, “Answering Queries Using Views,” *Proc. 14th ACM PODS*, pp. 95–104, 1995.

[14] L. Libkin, and L. Wong, “On Representation and Querying Incomplete Information in Databases with Bags,” *Information Processing Letters*, 56, pp. 209–214, 1995.

[15] L. Libkin, and L. Wong, “Query Languages for Bags and Aggregate Functions,” *J. Computer and System Sciences*, 55, pp. 241–272, 1997.

[16] R. Reiter, “On Closed World Databases,” *Logic and Databases, H. Gallaire and J. Minker (eds.)*, Plenum Press, pp. 55–76, 1978.

[17] A. Yamaguchi, Y. Ishihara, and T. Fujiwara, “Closure Properties of a Bag-based Data Model for Incomplete Information,” *DEWS2004*, http://www.ieice.org/iss/de/DEWS/proc/2004/program.html, 5-C-02, 2004.