GENERALIZED CONTAINMENT OF CONJUNCTIVE QUERIES

by

Yannis E. Ioannidis
Raghu Ramakrishnan

Computer Sciences Technical Report #1068

January 1992

Generalized Containment of Conjunctive Queries

Yannis E. Ioannidis*
Raghu Ramakrishnan†
Computer Sciences Department
University of Wisconsin
Madison, WI 53706
{yannis,raghu}@cs.wisc.edu

January 4, 1992

Abstract

Conjunctive queries are queries over a relational database, and are composed of the relational algebra operators *select*, *project* and *cartesian product*. In this paper, we study conjunctive queries over databases in which each tuple has an associated *label*; as a special case, in a traditional relational database, the label associated with a tuple is either 1 (meaning that the tuple is 'in' the relation) or 0 (meaning that the tuple is not in the relation). In particular, this generalized notion of a database allows us to consider relations that are fuzzy sets or multisets. Conjunctive queries over a relational database can be viewed as functions from sets to sets and containment (and equivalence) can be naturally defined based on set inclusion. It is known that the containment problem for conjunctive queries over a relational database is NP-complete. We examine this problem for databases in which tuples have associated labels, and establish results for a variety of *label systems*, that is, various (algebraic) conditions on the labels that can be associated with tuples.

1 Introduction

The problem of syntactically characterizing containment and equivalence of conjunctive queries was solved in the late 1970s by the work of Chandra and Merlin [CM77] and by the tableau work of Aho, Sagiv and Ullman [ASU79]. In both efforts, conjunctive queries were seen as functions from sets to sets and containment was naturally defined based on set inclusion.

It is sometimes necessary to go beyond the traditional model in which a given tuple is either in or not in a given relation, and to associate a *label* with each tuple [IW91]. As an example, a positive integer multiplicity (or number of copies, intuitively) is associated with every tuple in a multiset. Another example is fuzzy sets, where an arbitrary number in [0,1] (or certainty factor, intuitively) is associated with every tuple in a fuzzy set.

The results presented in this paper generalize earlier results on traditional conjunctive queries and also cover the example extensions mentioned above. The conditions under which these results are applicable are stated in terms of the algebraic properties of the set of labels and associated label

*Partially supported by the National Science Foundation under PYI Grant IRI-9157368 and by grants from DEC, HP, and AT&T.

†Partially supported by a David and Lucile Packard Foundation Fellowship in Science and Engineering, by the National Science Foundation under a Presidential Young Investigator Award and under grant IRI-9011563, and by grants from DEC, Tandem, and Xerox.

operations. Thus, they are also applicable to other possibly interesting extensions that have similar properties.

The paper is organized as follows. In Section 2, we present some background material and also develop our generalized notion of a database, which we consider to be one of our important contributions. We introduce label systems and identify two classes of label systems (types A and B) for which results are presented in this paper. In Section 3, we establish a very general necessary condition for conjunctive query containment that is satisfied by a large class of label systems, including type A and type B systems. In Section 4, we show that the necessary condition identified in the previous section is also sufficient for conjunctive query containment over databases with label systems of type A. This result strictly generalizes the condition of [CM77], and is also applicable to databases that deal with fuzzy sets. In Section 5, we present a sufficient condition for containment over databases with label systems of type B. For restricted classes of conjunctive queries in which there are no repeated predicates, we prove that this condition is necessary as well as sufficient. These results are applicable to databases that deal with multisets. We present our conclusions in Section 6.

2 Background and Basic Definitions

We review some standard concepts and develop our model of databases and conjunctive queries in this section. In particular, the definitions of label systems, databases and conjunctive query containment are generalizations of the usual definitions.

Definition 2.1 A *conjunctive query* is a first-order formula of the form $A_1 \wedge A_2 \wedge \ldots \wedge A_m \rightarrow C$. All the variables appearing in the formula are (implicitly) universally quantified. The formula to the left of $\rightarrow$ is called the *antecedent* and that to the right of $\rightarrow$ the *consequent*. Each one of $C, A_1, A_2, \ldots, A_m$ is an *atomic formula* of the form $Q(t_1, t_2, \ldots, t_n)$, where Q is a relation (predicate) symbol and t_i, $1 \leq i \leq n$, is a variable or a constant.

We use the convention that the atomic formula in the consequent of a conjunctive query always has the distinguished predicate symbol P, possibly subscripted with an indicator of the specific conjunctive query. Also, unless otherwise noted, the terms 'atomic formulas of a conjunctive query' or 'predicates of a conjunctive query' are used to refer to those in the antecedent.

Definition 2.2 A *label system* $\mathcal{L}$ is a quintuple $\mathcal{L} =< L, *, +, 0, \leq >$ such that:

- L is a domain of labels equipped with a partial order $\leq$.
- $*$ is a binary operation (called *multiplication*) on L that is associative and commutative.
- $+$ is a binary operation (called *addition*) on L that is associative and commutative.
- 0 is the additive identity in L and is also an annihilator with respect to multiplication and the least element with respect to the partial order $\leq$, i.e., $\forall a \in L, a + 0 = a, a * 0 = 0$, and $0 \leq a$.

When $a \leq b$ and $a \neq b$, then we use the notation $a < b$.

Definition 2.3 A label system $\mathcal{L} =< L, *, +, 0, \leq >$ is of *type A* if it satisfies the following:

(A1) $\forall a, b \in L - \{0\}, 0 < a * b \leq a$.

(A2) $\forall a \in L, a * a = a.$

(A3) $\forall a, a', b, b' \in L, (a \leq a' \text{ and } b \leq b') \Rightarrow a + b \leq a' + b'.$

(A4) $\forall a, b \in L, a + b \leq a \text{ or } a + b \leq b.$

Note that condition (A2) states that multiplication is idempotent.

Definition 2.4 A label system $\mathcal{L} =< L, *, +, 0, \leq >$ is of *type B* if it satisfies the following:

(B1) $\forall a, b \in L - \{0\}, a \leq a * b.$

(B2) $\forall a, a', b, b' \in L, (a \leq a' \text{ and } b \leq b') \Rightarrow a + b \leq a' + b'.$

(B3) $\forall a \in L, \exists a' \in L, a < a'.$

Definition 2.5 Let Q be an n-ary predicate symbol and $D_1, \ldots, D_n$ be the domains of values of the arguments of Q. Also let $\mathcal{L}$ be a label system with domain L. A *relation instance* for Q with respect to $\mathcal{L}$ is a total function [1] $\mathbf{Q}$: $D_1 \times \cdots \times D_n \rightarrow L$. Relation instances over the same cross product of domains are called *compatible*. A *database instance* with respect to a label system $\mathcal{L}$ is a set of relation instances. Any element of the domain $D_1 \times \ldots \times D_n$ is called a *tuple* and is denoted by $< d_1, \ldots, d_n >$, for $d_i \in D_i, 1 \leq i \leq n$.

In the sequel, whenever we refer to a database instance, it is understood that it is with respect to a given label system. Traditionally the set L is equal to {0,1} and a relation instance is compactly viewed as the subset of the cross product containing the tuples that map to 1.

Definition 2.6 Consider a conjunctive query α of the form $A_1 \wedge A_2 \wedge \ldots \wedge A_m \rightarrow C$. A *valuation* θ of α is a pair of functions $< \theta_v, \theta_l >$. Function θ_v is from the variables of α to some set of constants and function θ_l is from the atomic formulas of α to labels such that

$$\theta_l(C) = \prod_{i=1}^{m} \theta_l(A_i).$$

Applying θ on α gives an *instance* of α.

Definition 2.7 Consider two conjunctive queries α and β with compatible consequents whose distinguished variable in the i-th argument position, $1 \leq i \leq n$, is a_i and b_i, respectively. Let θ^α and θ^β be valuations of α and β, respectively. If for all $1 \leq i \leq n$, $\theta_v^\alpha(a_i) = \theta_v^\beta(b_i)$, then θ^α and θ^β are *compatible*.

Note that, in the above definition, α and β do not have to be distinct. For valuations of the same conjunctive query, it is easy to show that compatibility is an equivalence relation over valuations.

Definition 2.8 A valuation θ of a conjunctive query α is *true* with respect to a database instance if for every atomic formula $Q(x_1, \ldots, x_n)$ in α, $\mathbf{Q}(< \theta_v(x_1), \ldots, \theta_v(x_n) >) = \theta_l(Q(x_1, \ldots, x_n))$.

[1] Functions that denote relation instances appear in bold font.

Definition 2.9 Consider a conjunctive query α of the form $A_1 \wedge A_2 \wedge \ldots \wedge A_m \rightarrow C$ and a database instance I. Let Θ be the set of all valuations of α that are true with respect to I. Partition Θ based on the equivalence relation of valuation compatibility and let Θ_t denote the partition that generates tuple t in the distinguished variables of α. The *result* of applying α to I (denoted by $\alpha(I)$) is a relation instance $\mathbf{P}$, i.e., a function, such that

$$\mathbf{P}(t) = \sum_{\theta \in \Theta_t} \theta_I(C) = \sum_{\theta \in \Theta_t} \prod_{i=1}^{m} \theta_I(A_i).$$

Definition 2.10 Consider two conjunctive queries α and β with compatible consequents. A *homomorphism* $h : \beta \rightarrow \alpha$ is a total function from the variables of β into those of α, such that:

(i) If x, y are distinguished variables appearing in the same argument position in the consequent of β and α respectively, then $h(x) = y$.

(ii) If $Q(x_1, \ldots, x_n)$ appears in β, then $Q(h(x_1), \ldots, h(x_n))$ appears in α.

Note that a homomorphism $h : \beta \rightarrow \alpha$ induces a total function from the atomic formulas of β to the atomic formulas of α. Occasionally, when no confusion arises, we use h to denote that induced function as well.

Definition 2.11 For two functions f_1 and f_2 such that the range of f_2 is a subset of the domain of f_1, their *composition* is denoted by $f_1 \circ f_2$ and is defined as $(f_1 \circ f_2)(x) = f_1(f_2(x))$ for any member x in the domain of f_2.

Definition 2.12 Consider two relation instances $\mathbf{R}_1$ and $\mathbf{R}_2$ for a predicate R with domain $D_1 \times \cdots \times D_n$ with respect to a label system $\mathcal{L}$. $\mathbf{R}_1$ is *contained* in $\mathbf{R}_2$, denoted by $\mathbf{R}_1 \leq_r \mathbf{R}_2$, if for each tuple $t \in D_1 \times \cdots \times D_n$, $\mathbf{R}_1(t) \leq \mathbf{R}_2(t)$. Clearly, $\leq_r$ is a partial order.

Definition 2.13 For two conjunctive queries α and β, α is *more restrictive* than β, denoted $\alpha \leq_r \beta$, if for any database instance I, $\alpha(I) \leq_r \beta(I)$.

Note that the symbol $\leq_r$ is overloaded in that it signifies containment of relations as well as containment of conjunctive queries. This is natural, since the latter is defined in terms of the former. The ordering $\leq_r$ denotes a partial order over both the set of compatible relation instances and the set of conjunctive queries.

Definition 2.14 For two conjunctive queries α and β, α is *equivalent* to β, denoted $\alpha =_r \beta$, if $\alpha \leq_r \beta$ and $\beta \leq_r \alpha$.

3 Two General Results

In this section, we establish two results that are applicable to almost all label systems and are used extensively in the rest of the paper.

Lemma 3.1 Consider a label system $\mathcal{L}$ such that $\forall a, b \in L - \{0\}, a * b \neq 0$. For two conjunctive queries α and β, the inequality $\alpha \leq_r \beta$ holds with respect to $\mathcal{L}$ only if there exists a homomorphism $h : \beta \rightarrow \alpha$.

Proof: Let a_i (resp. b_i), $1 \leq i \leq n$, be the distinguished variable in the i-th argument position of α (resp. β). Assume that $\alpha \leq_r \beta$. Consider a valuation θ of α such that θ_v is one-to-one from the variables in α onto some set of constants C and θ_l maps all atomic formulas in α to nonzero elements of L. Consider a database instance such that for any relation $\mathbf{Q}$ the following is satisfied:

$$\mathbf{Q}(t) = \begin{cases} \theta_l(Q(x_1, \ldots, x_m)) & \text{if } t = < \theta_v(x_1), \ldots, \theta_v(x_m) > \text{ for some } Q(x_1, \ldots, x_m) \text{ in } \alpha \\ 0 & \textit{otherwise} \end{cases}$$

Let $\mathbf{P}_\alpha$ (resp. $\mathbf{P}_\beta$) be the result of applying α (resp. β) on that instance. Then, based on the requirements on $\mathcal{L}$ in the premise of the lemma, the following holds: $\mathbf{P}_\alpha(< \theta_v(a_1), \ldots, \theta_v(a_n) >) \neq 0$. Since $\alpha \leq_r \beta$ and because 0 is the least element of L with respect to $\leq$, it must be the case that $\mathbf{P}_\beta(< \theta_v(a_1), \ldots, \theta_v(a_n) >) \neq 0$ as well. Thus, a valuation θ' of β exists that is true with respect to the given database instance that is compatible with θ and such that for any atomic formula $Q(y_1, \ldots, y_m)$ in β, $\mathbf{Q}(< \theta'_v(y_1), \ldots, \theta'_v(y_m) >) \neq 0$. By the construction of the database instance, the above implies that θ'_v maps the variables of β into the set of constants C, Valuation θ_v is one-to-one and onto, so its inverse θ_v^{-1} is defined. Taking the composition $h = \theta_v^{-1} \circ \theta'_v$, it is easy to verify that it is a homomorphism from the variables of β to the variables of α. □

Lemma 3.2 Consider two conjunctive queries α and β and assume that there exists a homomorphism $h : \beta \rightarrow \alpha$. Consider a database instance I and the set Θ_α (resp. Θ_β) of all valuations of α (resp. β) that are true with respect to I. Let F be a total function on Θ_α ranging over the set of valuations of β and defined as $F(\theta) = \theta \circ h$. Function F has the following property:

For all $\theta \in \Theta_\alpha$, $F(\theta) \in \Theta_\beta$ and $F(\theta)$ is compatible with θ.

Proof: The proof of the lemma is based on the definition of homomorphisms. By property (ii) in Definition 2.10, for each atomic formula $Q(y_1, \ldots, y_k)$ in β, $Q(h(y_1), \ldots, h(y_k))$ appears in α as well. This implies the following:

$$\begin{aligned} \theta_l \circ h(Q(y_1, \ldots, y_m)) &= \theta_l(h(Q(y_1, \ldots, y_m))) \\ &= \theta_l(Q(h(y_1), \ldots, h(y_m))) \\ &= \mathbf{Q}(< \theta_v(h(y_1)), \ldots, \theta_v(h(y_m)) >) \text{ since } \theta \text{ is true with respect to } I \\ &= \mathbf{Q}(< \theta_v \circ h(y_1), \ldots, \theta_v \circ h(y_n)) >). \end{aligned}$$

Hence, by Definition 2.8 valuation $F(\theta) = \theta \circ h$ of β is true with respect to I.

Let a_i (resp. b_i), $1 \leq i \leq n$, be the distinguished variable in the i-th argument position of α (resp. β). By property (i) in Definition 2.10, $\theta_v(a_i) = \theta_v \circ h(b_i), 1 \leq i \leq n$, and therefore, θ and $F(\theta) = \theta \circ h$ are compatible. □

4 Label Systems of Type A

The following theorem identifies a necessary and sufficient condition for conjunctive query containment over databases with label systems of type A.

Theorem 4.1 For two conjunctive queries α and β, the inequality $\alpha \leq_r \beta$ holds with respect to a label system $\mathcal{L}$ of type A iff there exists a homomorphism $h : \beta \rightarrow \alpha$.

Proof: Let α be of the form $A_1 \wedge \ldots \wedge A_{m1} \rightarrow c_\alpha$ and β be of the form $B_1 \wedge \ldots \wedge B_{m2} \rightarrow c_\beta$. Also let a_i (resp. b_i), $1 \leq i \leq n$, be the distinguished variable in the i-th argument position of α (resp. β).

Assume that $\alpha \leq_r \beta$. By property (A1), it follows that $\forall a, b \in L - \{0\}, a * b \neq 0$. Therefore, by applying Lemma 3.1, the 'only-if' direction is proved.

For the 'if' direction, assume that there exists a homomorphism $h : \beta \rightarrow \alpha$. Consider a database instance I and the set Θ_α (resp. Θ_β) of all valuations of α (resp. β) that are true with respect to I. Let F be defined as in Lemma 3.2. Then, F has the following additional property:

(a) For all $\theta \in \Theta_\alpha$, $\theta_I(c_\alpha) \leq (F(\theta))_I(c_\beta)$, that is, $\theta_I(c_\alpha) \leq \theta_I \circ h(c_\beta)$.

The proof of property (a) is based on specific characteristics of label systems of type A. Let $A = \{A_i : 1 \leq i \leq m1\}$ and $B = \{B_i : 1 \leq i \leq m2\}$ represent the set of atomic formulas in α and β, respectively. Consider the subset A_B of A consisting of the atomic formulas in α that are images of atomic formulas in β under h. Without loss of generality, assume that $A_B = \{A_i : 1 \leq i \leq k\}$ for some $k \geq 1$. Thus, the following holds:

$$\theta_I \circ h(c_\beta) = \prod_{i=1}^{m2} \theta_I \circ h(B_i) = \prod_{i=1}^{k} \theta_I(A_i). \tag{1}$$

The last equality is due to property (A2), which implies that even if $h(B_i) = h(B_j)$ for some $i \neq j$, the product is not affected. On the other hand, the following is also true:

$$\begin{aligned} \theta_I(c_\alpha) &= \prod_{i=1}^{m1} \theta_I(A_i) = \prod_{i=1}^{k} \theta_I(A_i) * \prod_{i=k+1}^{m1} \theta_I(A_i) \\ &= \theta_I \circ h(c_\beta) * \prod_{i=k+1}^{m1} \theta_I(A_i) \text{ due to (1)} \\ &\leq \theta_I \circ h(c_\beta). \text{ by property (A1) of multiplication} \end{aligned}$$

From the above, we conclude that $\theta_I(c_\alpha) \leq \theta_I \circ h(c_\beta)$, and therefore that property (a) of F holds.

We proceed with the proof of the theorem. Partition Θ_α and Θ_β based on the equivalence relation of valuation compatibility and let $\Theta_{t\alpha}$ and $\Theta_{t\beta}$ be the corresponding partitions that generate tuple t in the distinguished variables of α or β. Clearly, there is a one-to-one and onto correspondence between the partitions obtained for α and those obtained for β (since for both of them, there is a single partition for each tuple in the domain of the consequent relation). Let V_α and V_β be multisets defined as follows: $V_\alpha = \{\theta_I(c_\alpha) : \theta \in \Theta_{t\alpha}\}$ and $V_\beta = \{\theta'_I(c_\beta) : \theta' \in \Theta_{t\beta}\}$. By property (A4) of addition, there is some element $v_0 \in V_\alpha$ such that

$$\sum_{v \in V_\alpha} v \leq v_0. \tag{2}$$

Suppose that $v_0 = \theta_I(c_\alpha)$ and $v'_0 = F(\theta_I)(c_\beta)$, for some $\theta \in \Theta_{t\alpha}$. From property (a) above, $v_0 \leq v'_0$, and from Lemma 3.2, $v'_0 \in V_\beta$. By property (A3) of addition, the following holds:

$$v_0 = v_0 + 0 \leq v'_0 + \sum_{v \in V_\beta - \{v'_0\}} v = \sum_{v \in V_\beta} v. \tag{3}$$

Combining (2) and (3) yields $\sum_{v \in V_\alpha} v \leq \sum_{v \in V_\beta} v$. If $\alpha(I)$ and $\beta(I)$ are equal to the relations $\mathbf{P}_\alpha$ and $\mathbf{P}_\beta$, respectively, by Definition 2.9, the above implies that for all tuples t in the result of α or

β, $\mathbf{P}_\alpha(t) \leq \mathbf{P}_\beta(t)$. Therefore, for an arbitrary database instance I, $\alpha(I) \leq_r \beta(I)$, which also implies that $\alpha \leq_r \beta$. □

Given the above theorem testing for conjunctive query containment with respect to type A label systems is identical to the same problem for the traditional relational databases. Hence, by the result of Chandra and Merlin [CM77], we have the following:

Proposition 4.1 Testing for conjunctive query containment with respect to type A label systems is NP-complete.

Example 4.1 Conjunctive queries over relational databases are a special case of type A systems. The only labels associated with tuples are 1 and 0, denoting that the tuple is in or not in a relation, respectively. The operation * is logical *and* and the operation + is logical *or*. The label 0 serves as the additive identity and annihilator for multiplication. It is easy to verify that all the conditions for a type A label system are satisfied. Thus, Theorem 4.1 generalizes the result of Chandra and Merlin [CM77]. □

Example 4.2 Another example of a type A system is a database in which every relation is a fuzzy set [Zad65]. The set of labels is the set of real numbers between 0 and 1, the operation * is *min* and the operation + is *max*. Again, 0 serves as the additive identity and multiplicative annihilator. All conditions for a type A system are satisfied. □

5 Label Systems of Type B

The following theorem identifies a sufficient condition for conjunctive query containment over databases with label systems of type B. Unfortunately, as Example 5.1 illustrates, it is not necessary in general.

Theorem 5.1 For two conjunctive queries α and β, if there exists an onto homomorphism $h : \beta \rightarrow \alpha$, then the inequality $\alpha \leq_r \beta$ holds with respect to a label system $\mathcal{L}$ of type B.

Proof: Let α be of the form $A_1 \wedge \ldots \wedge A_{m1} \rightarrow c_\alpha$ and β be of the form $B_1 \wedge \ldots \wedge B_{m2} \rightarrow c_\beta$. Also let a_i (resp. b_i), $1 \leq i \leq n$, be the distinguished variable in the i-th argument position of α (resp. β).

Assume that there exists an onto homomorphism $h : \beta \rightarrow \alpha$. Consider a database instance I and the set Θ_α (resp. Θ_β) of all valuations of α (β) that are true with respect to I. Let F be defined as in Lemma 3.2. Then, F has the following additional properties:

(a) For all $\theta \in \Theta_\alpha$, $\theta_l(c_\alpha) \leq (F(\theta))_l(c_\beta)$, that is, $\theta_l(c_\alpha) \leq \theta_l \circ h(c_\beta)$.

(b) F is one-to-one from Θ_α to Θ_β.

The proof of property (a) is based on the specific characteristics of label systems of type B. Let $A = \{A_i : 1 \leq i \leq m1\}$ and $B = \{B_i : 1 \leq i \leq m2\}$ represent the set of atomic formulas in α and β, respectively. Because h is onto, the set B can be partitioned into two subsets, say B_A and B_Q,

such that $h : B_A \rightarrow A$ is a bijection. Without loss of generality, assume that the two subsets are $B_A = \{B_i : 1 \leq i \leq m1\}$ and $B_Q = \{B_i : m1 + 1 \leq i \leq m2\}$. Thus, the following holds:

$$\begin{aligned}\theta_l \circ h(c_\beta) &= \prod_{i=1}^{m2} \theta_l \circ h(B_i) = \prod_{i=1}^{m1} \theta_l \circ h(B_i) * \prod_{i=m1+1}^{m2} \theta_l \circ h(B_i) \\ &= \prod_{i=1}^{m1} \theta_l(A_i) * \prod_{i=m1+1}^{m2} \theta_l \circ h(B_i) = \theta_l(c_\alpha) \prod_{i=m1+1}^{m2} \theta_l \circ h(B_i).\end{aligned}$$

From the above, because of property (B1) of multiplication, we conclude that $\theta_l(c_\alpha) \leq \theta_l \circ h(c_\beta)$, and therefore that property (a) of F holds.

The proof of property (b) consists of showing that, given two valuations $\theta, \theta' \in \Theta_\alpha$, if $\theta(\alpha) \neq \theta'(\alpha)$, then $\theta \circ h(\beta) \neq \theta' \circ h(\beta)$. Because $\theta(\alpha) \neq \theta'(\alpha)$, there must be at least one variable x in α such that

$$\theta_v(x) \neq \theta'_v(x). \tag{4}$$

Because h is onto, every variable of α is an h-image of some variable of β. Assume that $x = h(y)$, for some variable y of β. The above combined with (4) implies that

$$\theta_v \circ h(y) \neq \theta'_v \circ h(y). \tag{5}$$

Therefore, $\theta_v \circ h(\beta) \neq \theta'_v \circ h(\beta)$, which implies that property (c) of F holds.

We proceed with the proof of the theorem. Partition Θ_α and Θ_β based on the equivalence relation of valuation compatibility and let $\Theta_{t\alpha}$ and $\Theta_{t\beta}$ be the corresponding partitions that generate tuple t in the distinguished variables of α or β. As in the proof of Theorem 4.1, there is a one-to-one and onto correspondence between the partitions obtained for α and those obtained for β. Let V_α and V_β be multisets defined as follows: $V_\alpha = \{\theta_l(c_\alpha) : \theta \in \Theta_{t\alpha}\}$ and $V_\beta = \{\theta'_l(c_\beta) : \theta' \in \Theta_{t\beta}\}$. Let V_β^α be defined as follows: $V_\beta^\alpha = \{\theta'_l(c_\beta) : \theta' \in \Theta_{t\beta}$ and $\theta' = \theta \circ h$ for some $\theta \in \Theta_\alpha\}$. The properties of F imply that, for every element $v \in V_\alpha$ there is an element $v' \in V_\beta^\alpha$ that corresponds to v (Lemma 3.2) such that $v \leq v'$ (property (a)), which corresponds to no other element of V_α (property (b)). By property (B2) of addition, the above imply the following:

$$\sum_{v \in V_\alpha} v \leq \sum_{v' \in V_\beta^\alpha} v' \leq \sum_{v' \in V_\beta} v'. \tag{6}$$

If $\alpha(I)$ and $\beta(I)$ are equal to the relations $\mathbf{P}_\alpha$ and $\mathbf{P}_\beta$, respectively, by Definition 2.9, (6) implies that for all tuples t in the result of α or β, $\mathbf{P}_\alpha(t) \leq \mathbf{P}_\beta(t)$. Therefore, for an arbitrary database instance I, $\alpha(I) \leq_r \beta(I)$, which also implies that $\alpha \leq_r \beta$. □

The following proposition provides a straightforward necessary condition for conjunctive query containment with respect to label systems of type B.

Proposition 5.1 For two conjunctive queries α and β, the inequality $\alpha \leq_r \beta$ holds with respect to a label system $\mathcal{L}$ of type B only if there exists a homomorphism $h : \beta \rightarrow \alpha$.

Proof: Assume that $\alpha \leq_r \beta$. By property (B1), it follows that $\forall a, b \in L - \{0\}, a * b \neq 0$. Therefore, by applying Lemma 3.1, the proposition is proved. □

By restricting the form of conjunctive queries, the following theorem shows that the condition of Theorem 5.1 is both necessary and sufficient.

Theorem 5.2 For two conjunctive queries α and β such that α does not contain repeated predicates, the inequality $\alpha \leq_r \beta$ holds with respect to a label system $\mathcal{L}$ of type B iff there exists an onto homomorphism $h : \beta \rightarrow \alpha$.

Proof: The 'if' direction is an immediate consequence of Theorem 5.1. Suppose that $\alpha \leq_r \beta$. By Proposition 5.1, we know that there must be some homomorphism $h : \beta \rightarrow \alpha$. We first show that, in this case, there is a unique such homomorphism. Since there are no repeated predicates in α, for each atomic formula of β, there is a unique atomic formula in α that can be its image under any homomorphism. Therefore, for each variable in β its image is uniquely determined, i.e., there is a unique homomorphism $h : \beta \rightarrow \alpha$.

It remains to be shown that this unique homomorphism is onto. Assume to the contrary that h is not onto. Then, there is an atomic formula in α that is not the image of any atomic formula in β. Let Q be the predicate in that atomic formula of α. Clearly, Q cannot appear in β. Without loss of generality, assume that all arguments of all predicates in the conjunctive queries have the same domain. Consider a constant c in that domain and a database instance such that each relation $\mathbf{R}$ satisfies the following:

$$\begin{aligned} \mathbf{R}(t) &\neq 0 \qquad \text{if } t \text{ has } c \text{ in all its arguments} \\ \mathbf{R}(t) &= 0 \qquad \text{otherwise.} \end{aligned}$$

Let $\mathbf{P}_\alpha$ (resp. $\mathbf{P}_\beta$) be the result of applying α (resp. β) on that instance. Let s be the tuple in these results that has c in all of its arguments. Let $\mathbf{P}_\beta(s) = l$, where by the construction of the database instance, $l \in L - \{0\}$. By property (B3), there is a label $m \in L$ such that $m > l$. Suppose that t' is the tuple in Q that has c in all of its arguments and choose $\mathbf{Q}(t') = m$. Then, by property (B1), $\mathbf{P}_\alpha(t) \geq m > l = \mathbf{P}_\beta(t)$, which implies that $\alpha \not\leq_r \beta$, which is a contradiction. Hence, h must be onto. □

It is natural to ask whether Theorem 5.2 can be strengthened to cover the case that β does not contain repeated predicates (while α possibly does). Unfortunately, the following example shows that this is not possible.

Example 5.1 Consider the following two conjunctive queries:

$$\begin{aligned} \alpha &: Q(x) \wedge Q(x) \rightarrow P(x) \\ \beta &: Q(x) \rightarrow P(x) \end{aligned}$$

Let $\mathcal{L}$ be defined as follows: $L = N$ (the set of natural numbers including 0), $*$ is max, $+$ is the usual addition, and $\leq$ is the usual total order over the natural numbers. Clearly, $\alpha \leq_r \beta$ although there is no onto homomorphism from β to α. □

By limiting the definition of type B label systems so that examples like the above are excluded, we can prove a stronger version of Theorem 5.2.

Definition 5.1 A label system $\mathcal{L} =< L, *, +, 0, \leq >$ is of *type* B^- if it satisfies the following:

(B^-1) $\quad \forall a, b \in L - \{0\}, a \leq a * b$ and $\exists a \in L, \forall k \geq 1, a^k < a^{k+1}$.

(B^-2) $\quad \forall a, a', b, b' \in L, (a \leq a'$ and $b \leq b') \Rightarrow a + b \leq a' + b'$.

(B^-3) $\quad \forall a \in L, \exists a' \in L, a < a'$.

Observe that the only difference between label systems of type B and of type B$^-$ is that there is an element in L whose product with itself is strictly larger than the element itself (property (B$^-$1)). For these label systems, we have the following result.

Theorem 5.3 For two conjunctive queries α and β such that either α or β does not contain repeated predicates, the inequality $\alpha \leq_r \beta$ holds with respect to a label system $\mathcal{L}$ of type B$^-$ iff there exists an onto homomorphism $h : \beta \rightarrow \alpha$.

Proof: The 'if' direction as well the 'only-if' direction for the case where there are no repetitions in α are immediate consequences of Theorem 5.2. Suppose that β does not contain repeated predicates and that $\alpha \leq_r \beta$. By Proposition 5.1, we know that there must be some homomorphism $h : \beta \rightarrow \alpha$. Clearly, every such homomorphism must be one-to-one with respect to atomic formulas, since there is no repetition or predicates in β. Assume that h is not onto. Then, if m (resp. n) is the number of atomic formulas in α (resp. β), then clearly $m > n$.

Let l be an element of L such that $\forall k \geq 1, l^k < l^{k+1}$. Property (B$^-$1) ensures the existence of such a label. Without loss of generality, assume that all arguments of all predicates in the conjunctive queries have the same domain. Consider a constant c in that domain and a database instance such that each relation $\mathbf{R}$ satisfies the following:

$$\mathbf{R}(t) = l \qquad \text{if } t \text{ has } c \text{ in all its arguments}$$
$$\mathbf{R}(t) = 0 \qquad \text{otherwise.}$$

Let $\mathbf{P}_\alpha$ (resp. $\mathbf{P}_\beta$) be the result of applying α (resp. β) on that instance. Let s be the tuple in these results that has c in all of its arguments. We note that $\mathbf{P}_\alpha(s) = l^m$ and $\mathbf{P}_\beta(s) = l^n$. Since $m > n$, by property (B$^-$1), it follows that $\mathbf{P}_\alpha(s) > \mathbf{P}_\beta(s)$, which implies that $\alpha \not\leq_r \beta$, which is a contradiction. Hence, h must be onto. □

We remark that the choice of h in the above proof was arbitrary. Hence, $\alpha \leq_r \beta$ only if every homomorphism $h : \beta \rightarrow \alpha$ is onto.

Example 5.2 A database in which relations are multisets of tuples is an example of a type B system. The set of labels is the set of non-negative integers, the operation * is product and the operation + is sum. The number 0 is the additive identity and multiplicative annihilator. It is easy to verify that all the conditions for a type B label system are satisfied. □

6 Summary and Future Work

We have generalized the notion of a relational database to cover fuzzy sets, multisets, and other refinements to the concept of a relation as a set. We have examined the problem of conjunctive query containment for two important classes of systems. An interesting open problem is that of conjunctive query containment for type B systems. We presented a necessary and sufficient condition for queries with no repeated predicates and a sufficient condition for the general case. Is there a general necessary and sufficient condition? Other open problems include identifying other useful types of label systems and syntactically characterizing containment for them, and also possibly extending our results for unions and complements of conjunctive queries, generalizing the results of Sagiv and Yannakakis for the traditional case [SY80]. A more basic question is whether or not our conditions on label systems can be made more liberal. In particular, can we relax the requirement that the least element should be both the additive identity and the multiplicative annihilator?

References

[ASU79] A. Aho, Y. Sagiv, and J. Ullman. Equivalences among relational expressions. *SIAM Journal on Computing*, 8(2):218–246, May 1979.

[CM77] A. K. Chandra and P. M. Merlin. Optimal implementation of conjunctive queries in relational data bases. In *Proc. 9th Annual ACM Symposium on Theory of Computing*, pages 77–90, Boulder, CO, May 1977.

[IW91] Y. E. Ioannidis and E. Wong. Towards and algebraic theory of recursion. *JACM*, 38(2):329–381, April 1991.

[SY80] Y. Sagiv and M. Yannakakis. Equivalences among relational expressions with the union and difference operators. *JACM*, 27(4):633–655, October 1980.

[Zad65] L. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.