Weak preservation of multi-valued fusion

Antoon Bronselaer[1] **Guy De Tré**[2]

[1]Department of Telecommunications and Information Processing, Ghent University

Abstract

Fusion functions are important data integration tools that map a multiset of objects (i.e., the sources) onto a single object (i.e., the solution). Traditionally, it is assumed that the objects of interest are single-valued. In this paper, it will be assumed that each object has a multi-valued data structure, leading to a framework of second order fusion functions. An important class of such functions is called preservative and is characterized by the fact that one of the input objects is returned. A disadvantage of these functions is the a-priori limitation of the output space to the input objects. It is investigated here how this disadvantage can be mitigated by studying the principle of weak-preservation. The main idea is hereby that, instead of preserving one of the sources, a characteristic feature of one of the sources is preserved. Three such features will be studied: cardinality, k-cut and multiplicity distribution. It will then be shown how weak-preservation can be utilized in the design of second order fusion functions.

Keywords: Fusion, Multiset, Preservation

1. Introduction

A challenging problem in many modern data management systems is how to deal with duplicate data. Informally, the problem of duplicate data adheres to the fact that there can exist multiple descriptions of the same real world entity within one (or more) database(s). Usually, relieving a database from its duplicate data requires two distinct steps: *match* and *fuse*. In the 'match' step [1, 2, 3, 4], the basic problem is to compare two pieces of data (e.g., database records) and to decide whether they are duplicate or not. The goal of the 'match' step is thus to find all duplicate data in a database. In the 'fuse' step [5, 6, 7, 8, 9], data that have been identified as being duplicate, must be fused into one piece of data that is considered as an optimal real-world description. At the formal level, a collection (usually a multiset) of sources is given, where each of the sources describes the same real-world entity in one specific way. A fusion function is applied to these sources, resulting in a solution to the fusion problem.

Within the scope of this paper, we are interested in the specific case where the sources (and thus also the solution) have a *multi-valued* data structure. More specifically, it is assumed here that a source S is a multiset rather than a single-valued object. Although this assumption is rather uncommon in literature on fusion (see [7, 9] for overview papers), fusion of multi-valued sources occurs quite naturally in many modern database systems due to the existence of *relationships*:

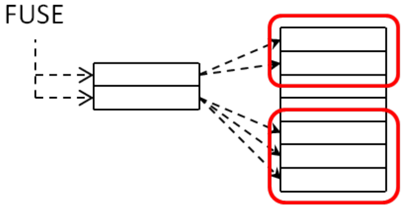

Figure 1: Multi-valued fusion in relational databases

- In relational database systems, a $1-n$ relationship implies that a tuple in a table can be linked to a *set of tuples* in another table as illustrated in Figure 1. As such, when duplicate tuples are fused in the leftmost table, the corresponding sets of tuples in the rightmost table should also be fused properly.
- In hierarchical database systems such as XML-databases, a similar reasoning holds. In Figure 2, two hierarchical data structures are shown that must be fused. This requires that the root nodes are fused first, followed by a proper fusion of the sets of child nodes of these roots.

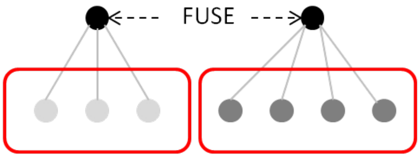

Figure 2: Multi-valued fusion in hierarchical databases

When designing fusion functions in general, an important role is played by *preservative* fusion functions (see [7, 10]). These functions are characterized by the fact that they select one of the input sources as the solution. Despite their advantages, a critique often attributed to these functions is that they are too restrictive in the sense that the a-priori solution space is limited to the sources. In order to mitigate this restrictiveness in the case of multi-valued data structures, the concept of *weak preservation* is introduced in this paper. Basically, this principle characterizes fusion functions for multi-valued data structures that preserve *features* of the sources, rather than the sources themselves. A simple example of such a feature is the cardinality. Indeed, when given three sources that all have cardinality 3, it seems reasonable to constrain the solution to a (multi)set with cardinality 3.

The remainder of this paper is structured as follows. In Section 2, the relevant related work regarding data fusion is discussed. In Section 3, some preliminary concepts involving multisets are presented. In Section 4, a simple framework of fusion functions is presented in which a distinction between first order and second order fusion functions is made. The concept of weak-preservation is studied in this framework and it is shown how this leads to a weak notion of idempotence. In Section 5, it is shown how weak-preservative fusion functions can be designed. Finally, in Section 6, the most important contributions of this paper are summarized.

2. Related work

In the past decades, many valuable contributions have been made to the field of data fusion. On the more formal level, Baral [11], Lin [12] and Konieczny [13] have investigated how fusion functions can be designed in the context of propositional belief bases modeled as a first-order theory. In this setting, information of different belief bases must be combined in a belief base that represents a maximal first-order theory. On a more practical level, Bleiholder [14] has proposed an extension of the standard SQL syntax to support redundancy removal operations, leading to the FUSE BY-operator. Bilke [6] and Naumann [7] have proposed the HumMer system, which is an integrated system that allows the semi-automatic integration of several heterogeneous data sources. Benjelloun [15] investigated the properties of match functions and fusion functions simultaneously and showed that whenever the match and fusion functions satisfy certain properties, very efficient fusion algorithms can be constructed. Another interesting approach is that of Motro [8], who models a fusion function as a multi-dimensional optimization problem. It is noted and emphasized that all of these approaches deal with the single-valued case. The multi-valued case has been studied to much lesser extent [16, 17], thus leaving an opportunity for further research. For additional readings on data fusion, the authors refer to [9] for a good and complete overview.

3. Preliminaries

A well known extension of Cantorian sets is that of multisets (also called bags) [18]. For many decades, these data structures have been of particular interest within the broad field of informatics. Within the scope of this paper, the notations of Yager are adopted [18].

Definition 1 (Multiset)
A multiset A over a universe U is defined by a function:

$$A : U \to \mathbb{N}.$$

For each $u \in U$, $A(u)$ denotes the multiplicity of u in A. The set of all multisets drawn from a universe U is denoted $\mathcal{M}(U)$.

The j-cut of a multiset A is a regular set, denoted as A_j and is given by:

$$A_j = \{u | u \in U \wedge A(u) \geq j\}.$$

Whenever we wish to assign an index $i \in \mathbb{N}$ to a multiset A, we use the notation $A_{(i)}$, while the notation A_j is reserved for the j-cut of A. The following operators on multisets are considered:

$$\begin{aligned} \forall u \in U: \quad & (A \cup B)(u) = && \max(A(u), B(u)) \\ \forall u \in U: \quad & (A \cap B)(u) = && \min(A(u), B(u)) \\ \forall u \in U: \quad & (A \oplus B)(u) = && A(u) + B(u). \end{aligned}$$

The relation $\subseteq$ is extended as follows:

$$A \subseteq B \Leftrightarrow (\forall u \in U : A(u) \leq B(u))$$

Similarly, the relation $\subset$ is extended as:

$$A \subset B \Leftrightarrow (A \subseteq B) \wedge (\exists u \in U : A(u) < B(u)).$$

The cardinality of a multiset A is calculated as the sum of all multiplicities:

$$|A| = \sum_{u \in U} A(u).$$

Finally, the relation $\in$ is extended as follows:

$$\forall A \in \mathcal{M}(U) : \forall u \in U : u \in A \Leftrightarrow A(u) \neq 0.$$

4. Fusion and Weak-Preservation

At the basis of this paper lies a simple framework of fusion in which distinction is made between first and second order fusion functions.

Definition 2 (First Order Fusion Function)
A first order fusion function over a universe U is defined by:

$$\mathrm{F} : \mathcal{M}(U) \to U.$$

Basically, a first order fusion function maps a multiset of elements from U onto an element in U. Hence, the sources and the solution are elements in U. Usually, it is assumed that elements in U have a single-valued data structure. For example, if $U = \mathbb{N}$, then min and max are two first order fusion functions over U. Many first order fusion functions described in literature satisfy a property called *preservation*.

Definition 3 (Preservation of F)
A first order fusion function F over a universe U is preservative if:

$$\forall M \in \mathcal{M}(U) : \mathrm{F}(M) \in M.$$

In the scope of this paper, the framework of fusion functions is extended to second order fusion functions, defined as follows.

Definition 4 (Second Order Fusion Function) *A second order fusion function over a universe U is defined by:*

$$\mathrm{F}^* : \mathcal{M}\big(\mathcal{M}(U)\big) \to \mathcal{M}(U).$$

In the case of second order fusion functions, the sources and the solution are now *multisets* of elements in U, rather than elements in U. If $U = \mathbb{N}$, then $\oplus$ (i.e., the multiset sum) is an example of a second order fusion function over U. For the sake of simplicity, second order fusion functions will be called 'fusion functions' for short in the remainder of this paper. An interesting class of fusion functions relies on the principle of *boundedness.*

Definition 5 (Boundedness) *A second order fusion function* F^* *over* U *is* bounded *if:*

$$\forall M \in \mathcal{M}\big(\mathcal{M}(U)\big) : \underline{\mathrm{F}}^*(M) \subseteq \mathrm{F}^*(M) \subseteq \overline{\mathrm{F}}^*(M)$$

where:

$$\begin{aligned} \underline{\mathrm{F}}^*(M) &\triangleq \bigcap_{S \in M} S \\ \overline{\mathrm{F}}^*(M) &\triangleq \bigcup_{S \in M} S. \end{aligned}$$

Boundedness implies that a solution consists of elements that occur in at least one of the sources. For that reason, it is usual a mandatory constraint for a second order fusion function in a practical setting. In the remainder of this paper, we shall assume that F^* is bounded, unless explicitly stated otherwise. Similar to the case of first order fusion functions, *preservative* fusion functions are defined as follows.

Definition 6 (Preservation of F^*) *A second order fusion function* F^* *over a universe* U *is preservative if:*

$$\forall M \in \mathcal{M}\big(\mathcal{M}(U)\big) : \mathrm{F}^*(M) \in M.$$

Preservative fusion functions are characterized by the fact that they select one of the input sources (i.e., they *preserve* one of the sources as the solution). It can be shown easily that preservative functions are always bounded, which makes the class of preservative functions a subclass of the class of bounded functions. Preservative fusion functions are a popular class of fusion functions (also in the first order case) because of their simplicity in definition. However, a critique is that they are too restrictive, because $\mathrm{F}^*(M)$ has only $|M_1|$ possible outcomes. Moreover, in the first order case, preservative fusion functions are typically based on some order relation over U (e.g., min and max). Finding such an order relation in the multi-valued case is much more difficult. In this paper it is proposed to *weaken* the principle of preservation. More specifically, instead of providing a solution that is equal to a source, it is proposed to provide a solution that inherits some important features of a source. As such, a class of fusion functions is obtained that does not preserve one of the sources, but rather preserves some features of one of the sources. This idea is formalized by introducing θ-preservation.

Definition 7 (θ-Preservation) *A second order fusion function* F^* *over a universe* U *is θ-preservative if:*

$$\forall M \in \mathcal{M}\big(\mathcal{M}(U)\big) : \exists S \in M : \theta\left(\mathrm{F}^*(M)\right) = \theta\left(S\right)$$

where θ is a mapping defined by $\theta : \mathcal{M}(U) \to \mathcal{X}$ and $\mathcal{X}$ is a feature space.

It can be seen that θ-preservation is a weaker variant of preservation. Indeed, if θ is the identity function, then θ-preservation is equivalent to preservation. This result shows that θ-preservation is a more general notion than preservation. An interesting observation regarding θ-preservation is that it provides a weakened notion of *idempotence.* Recall that idempotence of F^* means that if all sources in M are equal, then $\mathrm{F}^*(M)$ must be equal to this one source (Figure 3). It can be easily verified that preservative functions are always idempotent. In the case of θ-preservation, idempotence is weakened in the sense that, if sources in M are *equivalent* w.r.t. θ (i.e., they are not equal, but their image of θ is equal), then the result of $\mathrm{F}^*(M)$ must be equivalent with the sources. This is illustrated schematically in Figure 3. The notion of weak-idempotence justifies to a large extent why the concept of θ-preservation is a useful generalization of preservation. Indeed, if all

IDEMPOTENCE

$$\begin{tikzcd} M = \{S, ..., S\} \arrow[r, "\mathrm{F}^*"] & S \end{tikzcd}$$

θ-IDEMPOTENCE

$$\begin{tikzcd} M = \{S_1, ..., S_n\} \arrow[r, "\mathrm{F}^*"] \arrow[d, "\theta"] & \mathrm{F}^*(M) \arrow[d, "\theta"] \\ \theta(S_1) = ... = \theta(S_n) = C & \theta(\mathrm{F}^*(M)) = C \end{tikzcd}$$

Figure 3: Idempotence (upper panel) and weak idempotence (lower panel)

the sources are equivalent with respect to a certain feature, it can be strongly advocated that the solution must also satisfy this feature. In the following, three variants of θ-preservation will be studied: *cardinality* preservation, *k-cut* preservation and *multiplicity* preservation.

4.1. Cardinality Preservation

Perhaps the most obvious feature that qualifies for θ-preservation is cardinality. This type of θ-preservation is characterized by:

$$\theta(.) \triangleq |.|$$

and implies that the solution of fusion should be a multiset with cardinality equal to the cardinality of one of the sources. The feature space is given by $\mathcal{X} = \mathbb{N}$. Weak idempotence dictates in this case that if all sources have the same cardinality, the solution should be a multiset with this specific cardinality. Despite its intuitive acceptability, cardinality preservation is not a trivial property. It can for example be seen that $\underline{\mathrm{F}}^*$ (intersection) and $\overline{\mathrm{F}}^*$ (union) are not cardinality preservative.

Example 1
Let us consider $U = \{a, b, c, d\}$ and let $M = \{S_{(1)}, S_{(2)}, S_{(3)}\}$ with:

$$\begin{aligned} S_{(1)} &= \{a, a, a\} \\ S_{(2)} &= \{a, a, b\} \\ S_{(3)} &= \{d, c\} \end{aligned}$$

then cardinality preservation of F^* implies that $|\,\mathrm{F}^*(M)|$ is either 2 or 3.

With respect to the possessed properties, it can be seen that cardinality preservative fusion functions are not bounded, not idempotent and not preservative. It immediately follows that cardinality preservation is a very broad concept and is best paired with additional requirements on the definition of F^*. This principle will be elaborated further in Section 5.

4.2. k-Cut Preservation

A second interesting feature to preserve is the k-cut of one of the sources. In this case, θ-preservation is characterized by:

$$\theta(.) \triangleq ._k$$

Weak idempotence dictates in this case that if all sources have the same k-cut, then the solution should be a multiset with the same k-cut. A particularity in this case is that θ-preservation in fact adheres to a *set* of constraints on the solution (i.e., one constraint for each value of k). The size of this set is equal to the number of cuts that must be preserved. At the extremes there are two interesting cases:

- In one extreme, only one cut must be preserved. For example, if it is required that a fusion function must be 1-cut preservative, then the solution must contain the same elements as one of the sources. This constraint is of particular interest if the actual multiplicities of elements have no relevance. In the remainder, the case where only 1-cut preservation is required, will be referred to as *single-cut* preservation.
- In the other extreme, *all* cuts must be preserved. This means that the fusion function must be 1-cut preservative, 2-cut preservative,... In the remainder, the case where k-cut preservation is required for all values of k, will be referred to as *full-cut* preservation.

Note that if a fusion function must be k-cut preservative for multiple values of k, the corresponding rules are evaluated independently. To clarify this, consider $M = \{S_{(1)}, S_{(2)}\}$ and assume that $\mathrm{F}^*(M)_1 = \left(S_{(1)}\right)_1$ and $\mathrm{F}^*(M)_2 = \left(S_{(2)}\right)_2$, then F^* is both 1-cut and 2-cut preservative, even though the preserved cuts correspond to different sources. This is of course only possible if $\left(S_{(2)}\right)_2 \subseteq \left(S_{(1)}\right)_1$.

Example 2
Let us reconsider the sources from Example 1 and consider the following two fusion functions:

$$\begin{aligned} \mathrm{F}_1^*(M) &= \{d, d, c\} \\ \mathrm{F}_2^*(M) &= \{a, a, a, b\}. \end{aligned}$$

Both of these fusion functions are 1-cut preservative, because the 1-cut of $\mathrm{F}_1^*(M)$ equals the 1-cut of $S_{(3)}$, while the 1-cut of $\mathrm{F}_2^*(M)$ the 1-cut of $S_{(2)}$. While $\mathrm{F}_2^*(M)$ is 2-cut preservative (see $S_{(2)}$), $\mathrm{F}_1^*(M)$ is clearly not 2-cut preservative because there is no source with 2-cut equal to $\{d\}$. Finally, both fusion functions are 3-cut preservative. Fusion function F_1^* is an illustration of the fact that k-cut preservation is not a monotone feature in terms of k.

It can be seen that the class of k-cut preservative fusion functions is in general not idempotent, not bounded and not preservative. However, the following result can be shown.

Theorem 1
Let F^* be a full-cut preservative second order fusion function over U, then F^* is bounded.

Proof 1
A second order fusion function F^* is bounded if $\underline{\mathrm{F}}^*(M) \subseteq \mathrm{F}^*(M) \subseteq \overline{\mathrm{F}}^*(M)$.

(a) ***Proof of*** *$\underline{\mathrm{F}}^*(M) \subseteq \mathrm{F}^*(M)$. Suppose that $u \in \underline{\mathrm{F}}^*(M)_k$, for some $k \in \mathbb{N}$ and for some $M \in \mathcal{M}(\mathcal{M}(U))$. Taking into account the definition of multiset intersection (Section 3), it follows that all sources in M must have u in their k-cut. This means that:*

$$\forall S \in M : u \in S_k.$$

Now suppose that $u \notin \mathrm{F}^*(M)_k$, then we have that:

$$\forall S \in M : S_k \neq \mathrm{F}^*(M)_k$$

which is in contradiction with the fact that F^* is full-cut preservative and thus also k-cut preservative. In other words, full-cut preservation of F^* implies that:

$$u \in \underline{\mathrm{F}}^*(M)_k \Rightarrow u \in \mathrm{F}^*(M)_k$$

for any k. This means that:

$$\forall k \in \mathbb{N} : \underline{\mathrm{F}}^*(M)_k \subseteq \mathrm{F}^*(M)_k$$

which concludes the proof of the first part.

(b) ***Proof of*** $\mathrm{F}^*(M) \subseteq \overline{\mathrm{F}}^*(M)$. *Suppose that $\mathrm{F}^*(M)$ would not be a subset of $\overline{\mathrm{F}}^*(M)$, then there exists an element $u \in U$ such that:*

$$\mathrm{F}^*(M)(u) > \overline{\mathrm{F}}^*(M)(u) = \max_{S \in M} S(u).$$

Now let $k = \mathrm{F}^*(M)(u)$, then we have that:

$$u \in \mathrm{F}^*(M)_k$$

and

$$\forall S \in M : u \notin S_k.$$

As a consequence, the k-cut of $\mathrm{F}^*(M)$ can not be equal to the k-cut of any of the sources, which is in contrast with the fact that F^* is full-cut preservative. It thus follows that $\mathrm{F}^*(M) \subseteq \overline{\mathrm{F}}^*(M)$. □

Due to the fact that full-cut preservation implies boundedness, it also implies idempotence. It can be seen that $\underline{\mathrm{F}}^*$ and $\overline{\mathrm{F}}^*$ are not k-cut preservative.

4.3. Multiplicity Preservation

A third feature that is taken into account here is the multiplicity distribution of a multiset. Informally, given a multiset A, the multiplicity distribution is a multiset of natural numbers that indicates how many elements occur in A with a given multiplicity. The multiplicity distribution of a multiset is formalized as follows.

Definition 8 (Multiplicity Distribution)
Given a universe U, the multiplicity distribution of multisets over U is defined by:

$$\delta : \mathcal{M}(U) \rightarrow \mathcal{M}(\mathbb{N}_0)$$

such that for any $A \in \mathcal{M}(U)$:

$$\forall n \in \mathbb{N} : \delta(A)(n) = \left|A_n \ominus A_{n+1}\right|$$

where $\ominus$ is the set difference operator.

Example 3
Let us consider source $S_{(2)} = \{a, a, b\}$ from Example 1, then we have that:

$$\begin{aligned} \delta\left(S_{(2)}\right)(1) &= \left|\{a,b\} \ominus \{a\}\right| = 1 \\ \delta\left(S_{(2)}\right)(2) &= \left|\{a\} \ominus \emptyset\right| = 1. \end{aligned}$$

As such, the multiplicity distribution of multiset $S_{(2)}$ is a multiset $\delta\left(S_{(2)}\right) = \{1, 2\}$ which reflects that $S_{(2)}$ has one element with multiplicity 1 (i.e., b) and one element with multiplicity 2 (i.e., a).

Multiplicity preservation can now be characterized by:

$$\theta\,(.) \overset{\triangle}{=} \delta\,(.)$$

where the feature space is given by $\mathcal{X} = \mathcal{M}(\mathbb{N}_0)$. Weak idempotence dictates here that if all sources have an identical multiplicity distribution, then the solution should be a multiset with this same multiplicity distribution. It can be seen that the functions $\underline{\mathrm{F}}^*$ and $\overline{\mathrm{F}}^*$ are not multiplicity preservative. With respect to the basic properties, the class of multiplicity preservative functions is not bounded, not idempotent and not preservative. Again, this class of fusion functions is a very broad one that typically needs to be narrowed down with additional constraints on the definition of F^* (Section 5).

5. Construction of θ-preservative F^*

In the previous section, the concept of θ-preservation has been introduced and three instantiations of θ have been discussed. It was pointed out that the corresponding classes of θ-preservative fusion functions are very broad and, in their generality, fail to possess some interesting properties like boundedness and idempotence (see Table 1). This however does not render the principle of θ-preservation useless, as will be shown in this section. It will be pointed out that the power of θ-preservation lies in a smart choice of the source for which θ is preserved, especially when combining several types of θ-preservation. Next, a general framework for the design of θ-preservative fusion functions will be provided by modeling fusion as an optimization problem.

$\theta(.)$	Idempotent	Bounded	Preservative
$\lvert.\rvert$	-	-	-
$\cdot_k$	Full-cut	Full-cut	-
$\delta(.)$	-	-	-

Table 1: Summary of properties for θ-preservative second order fusion functions

5.1. Source selection

When dealing with the case of classical preservation (Definition 6), an important question is how to choose the source that is preserved. In the case of first order fusion, there is usually an order relation underlying the choice. Indeed, min and max are two well-known preservative first order fusion functions that are based on the idea of a total ordering over U (e.g., recency, reliability...). However, in the case of second order fusion, the presence of a relevant total order relation over $\mathcal{M}(U)$ is far less trivial. This gives preservative second order fusion functions an arbitrary nature, especially in the case when the principle of majority voting can not be used (i.e., because each source is unique). At this point, (combinations of) θ-preservation rules can lead to a useful and relevant solution by relying on a majority rule in the feature space $\mathcal{X}$. More specifically, instead of requiring that $\mathrm{F}^*(M)$ preserves the feature θ of an arbitrary source, it is possible to require that this feature must be possessed by as many sources as possible. This leads us to the principle of an $\mathcal{X}$-majority.

Definition 9 ($\mathcal{X}$-majority)
Let F^* be a second order fusion function over a universe U that is θ-preservative with $\theta : \mathcal{M}(U) \to \mathcal{X}$, then $\mathrm{F}^*(M)$ has an $\mathcal{X}$-majority if and only if:

$$\theta\left(\mathrm{F}^*\left(M\right)\right) = \arg\max_{x \in \mathcal{X}} \left| \{S | S \in M \wedge \theta(S) = x\} \right|.$$

Informally, $\mathrm{F}^*(M)$ has an $\mathcal{X}$-majority if and only if a majority of the sources shares the feature $\theta(.)$ with $\mathrm{F}^*(M)$.

Example 4
Let us consider $U = \{a, b, c\}$ and let $M = \{S_{(1)}, S_{(2)}, S_{(3)}, S_{(4)}\}$ with:

$$\begin{aligned} S_{(1)} &= \{a, a, b\} \\ S_{(2)} &= \{c, c, b\} \\ S_{(3)} &= \{b, b, b, b\} \\ S_{(4)} &= \{c, c, a\}. \end{aligned}$$

Assume a fusion function F^* that is required to be cardinality preservative. This means that $\mathrm{F}^*(M)$ must have cardinality 3 or 4. However, it can be seen that a solution with cardinality 3 will have an $\mathcal{X}$-majority because there are three sources with cardinality 3. Suppose that we now add multiplicity preservation with an $\mathcal{X}$-majority to the list of constraints, then we find that our solution must have a multiplicity distribution $\{2, 1\}$, meaning that there should be one element occurring once and one element occurring twice.

S	S_1	S_2
$S_{(1)}$	$\{a, b\}$	$\{a\}$
$S_{(2)}$	$\{b, c\}$	$\{c\}$
$S_{(3)}$	$\{b\}$	$\{b\}$
$S_{(4)}$	$\{a, c\}$	$\{c\}$

Table 2: 1-cuts and 2-cuts for Example 4

In order to specify the actual elements, let us also require 1-cut and 2-cut preservation. Table 2 shows the 1-cuts and 2-cuts of the considered sources. With respect to 2-cut preservation, we see that an $\mathcal{X}$-majority is obtained when the 2-cut equals $\{c\}$. With respect to the 1-cut, there is no $\mathcal{X}$-majority, but the choice of $\{c\}$ as 2-cut limits us to 1-cut $\{b, c\}$ or $\{a, c\}$. At this point, only two solutions still satisfy all the constraints that we have specified, that is:

$$\mathrm{F}^*(M) = \{a, c, c\} \vee \mathrm{F}^*(M) = \{b, c, c\}.$$

In this case, F^* chooses one of the sources, making it preservative. However, the choice made here is certainly not arbitrary as our specifications have excluded two sources.

Example 4 shows us that combining θ-preservation rules under the demand of $\mathcal{X}$-majority can lead us to relevant solutions. An important question is however how the reasoning followed in Example 4 can be used in the design of fusion functions.

5.2. θ-preservation and optimization

In this section, it is shown how θ-preservative fusion functions can be designed. One way of doing so, is modeling the problem of fusion as an optimization problem. Let us therefore consider an objective function $\mathcal{O}$ defined by:

$$\mathcal{O} : \mathcal{M}(U) \times \mathcal{M}(\mathcal{M}(U)) \to [0, 1]$$

such that $\mathcal{O}(A, M)$ evaluates the quality of A as a solution for sources M. Hereby, $\mathcal{O}(A, M) = 1$ means that A is a perfect solution and $\mathcal{O}(A, M) = 0$ means that A is completely rejected as a solution. We can then design a second order fusion function as follows:

$$\mathrm{F}^*(M) = \arg\max_{A \in \mathcal{M}(U)} \mathcal{O}(A, M).$$

There are several ways in which the objective function can be chosen. A very simple objective function is the mean Jaccard similarity:

$$\mathcal{O}(A, M) = \frac{1}{|M|} \sum_{S \in M} J(A, S) \tag{1}$$

where J is defined by:

$$J(A, S) = \frac{|A \cap S|}{|A \cup S|}.$$

More advanced alternatives can be found in [17] and [8]. Regardless of how the objective function is chosen, the framework of optimal fusion functions allows us to easily design θ-preservative fusion functions by limiting the search space to θ-preservative solutions as follows:

$$\mathrm{F}^*(M) = \arg\max_{A \in \mathcal{M}(U) \wedge (\exists S \in M : \theta(A) = \theta(S))} \mathcal{O}(A, M). \tag{2}$$

Equation (2) defines a function F^* that maximizes the objective function $\mathcal{O}$ under the constraint that the solution must preserve the feature θ. Needless to say, the principle of preserving the feature with an $\mathcal{X}$-majority can be easily added to specification of F^*.

Example 5
Let us reconsider the sources from Example 4. If we apply F^* *by using the objective function given in Equation* (1)*, then we find that* $\mathrm{F}^*_1(M) = \{b, c, c, a\}$*, which is cardinality preserving, but not with an* $\mathcal{X}$*-majority. If we add the constraint of cardinality preservation with* $\mathcal{X}$*-majority to the definition of* F^**, we find that* $\mathrm{F}^*_2(M) = \{b, c, c\}$*. Moreover, if we calculate the values of the objective function, we find that:*

$$\begin{aligned}\mathcal{O}\big(\mathrm{F}^*_1(M), M\big) &= 0.51\\ \mathcal{O}\big(\mathrm{F}^*_2(M), M\big) &= 0.47.\end{aligned}$$

This is a rather small difference in average similarity, compensated by the fact that F^*_2 *offers us a more intuitive solution w.r.t. its cardinality.*

6. Conclusion

In this paper, the concept of θ-preservation of second order fusion functions (i.e., fusion functions that operate on multi-valued sources) has been introduced. Whereas traditional preservative functions choose one of the sources (i.e., one of the sources is preserved), θ-preservative fusion functions preserve the feature θ of one of the sources. Three such features have been proposed: the cardinality, the k-cut(s) and the multiplicity distribution. It has been shown how θ-preservation implies a weaker notion of idempotence. More specifically, it has been shown how θ-preservative fusion functions are idempotent in the feature space $\mathcal{X}$. Next, it was pointed out that θ-preservative fusion functions can be designed by defining fusion as an optimization problem. In that setting, θ-preservation can be added as a boundary constraint. Finally, the principle of majority voting was introduced in the framework of θ-preservation.

References

[1] Howard Newcombe, James Kennedy, S Axford, and A James. Automatic linkage of vital records. *Science*, 130(3381):954–959, 1959.

[2] Ivan Fellegi and Alan Sunter. A theory for record linkage. *American Statistical Association Journal*, 64(328):1183–1210, 1969.

[3] Matthew Jaro. Advances in record linking methodology as applied to the 1985 census of tampa florida. *Journal of the American Statistical Society*, 84(406):414–420, 1989.

[4] Antoon Bronselaer, Axel Hallez, and Guy De Tré. Extensions of fuzzy measures and the sugeno integral for possibilistic truth values. *International Journal of Intelligent Systems*, 24(2):97–117, 2009.

[5] Isabelle Bloch. Information combination operators for data fusion: A comparative review with classification. *IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans*, 26(1):52–67, 1996.

[6] Alexander Bilke, Jens Bleiholder, Christoph Böhm, Karsten Draba, Felix Naumann, and Melanie Weis. Automatic data fusion with hummer. In *VLDB*, pages 1251–1254, 2005.

[7] Felix Naumann, Alexander Bilke, Jens Bleiholder, and Melanie Weis. Data fusion in three steps: Resolving inconsistencies at schema-, tuple-, and value-level. In *Bulletin of the technical committee on data engineering*, pages 21–31, 2006.

[8] Amihai Motro and Philipp Anokhin. Fusionplex: resolution of data inconsistencies in the integration of heterogeneous information sources. *Information Fusion*, 7(2):176–196, 2006.

[9] Jens Bleiholder and Felix Naumann. Data fusion. *ACM Computing Surveys*, 41(1), 2008.

[10] Antoon Bronselaer and Guy De Tré. Aspects of object merging. In *Proceedings of the NAFIPS Conference*, pages 27–32, Toronto, Canada, 07/2010 2010.

[11] Chitta Baral, Sarit Kraus, Jack Minker, and V Subrahmanian. Combining knowledge bases consisting of first-order theories. *Computational Intelligence*, 8(1):45–71, 1992.

[12] Jinxin Lin and Alberto Mendelzon. Merging databases under constraints. *International Journal of Cooperative Information Systems*, 7(1):55–76, 1998.

[13] Sébastien Konieczny and Ramon Pérez. Merging information under constraints: a logical framework. *Journal of Logic and Computation*, 12(1):111–120, 2002.

[14] Jens Bleiholder, Melanie Herschel, and Felix Naumann. Eliminating nulls with subsumption and complementation. *IEEE Data Engineering Bulletin*, 34(3):18–25, 2011.

[15] Omar Benjelloun, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, and Jennifer Widom. Swoosh: a generic approach to entity resolution. *VLDB Journal*, 18(1):255–276, 2009.

[16] Nicholas Rescher and Ruth Manor. On inferences from inconsistent premises. *Theory and Decision*, 1:179–217, 1970.

[17] Antoon Bronselaer, Daan Van Britsom, and Guy De Tré. A framework for multiset merging. *Fuzzy Sets and Systems*, 191:1–20, 2012.

[18] Ronald Yager. On the theory of bags. *International Journal of General Systems*, 13(1):23–27, 1986.