Aspects of object merging

Antoon Bronselaer
Department of Telecommunications
and Information Processing
Ghent University
Sint-Pietersnieuwstraat 41, B9000, Ghent, Belgium
Email: antoon.bronselaer@ugent.be

Guy De Tré
Department of Telecommunications
and Information Processing
Ghent University
Sint-Pietersnieuwstraat 41, B9000, Ghent, Belgium
Email: guy.detre@ugent.be

***Abstract*—Information fusion is a research area that investigates how to combine information provided by independent sources into one piece of information. This topic has been studied for several applications leading to, amongst others, aggregation operators in bounded lattices and merge functions of propositional belief bases. In this paper, information fusion is investigated in the context of coreferent objects, which are objects that refer to the same real world entity. Some important properties of object merge functions are pointed out and object merge functions for both atomic and complex objects are investigated in a possibilistic framework. It is shown how merge functions for complex objects can be composed of merge functions for atomic objects, such that the composite function inherits the properties of the merge functions from which it is composed.**

I. Introduction

Information fusion is a research area that deals with the combination of information provided by independent sources into one piece of information. The challenge hereby is to resolve inconsistencies between the different sources. An interesting aspect of information fusion is its applicability in many different contexts. In a mathematical context, information fusion has led to the development of numerous aggregation operators such as triangular norms and conorms [1], generalized means [2], [3] and uninorms [4]. Aggregation operators fuse information that is represented as an element of a complete lattice $(L, \leq)$. The information typically expresses *facts*, for example the opinion or score of an agent. Next to aggregation operators, a significant body of research deals with the case where each source is considered to be a *propositional belief base* modeled as a first-order theory [5], [6], [7], [8], [9], [10]. A typical difference between propositional belief bases and aggregation operators, is the presence of non-factual knowledge, such as inference rules and integrity constraints. As a consequence, the interest here is to *combine* all information in a maximal first-order theory. Such a setting occurs, amongst others, in heterogeneous databases [11]. A third type of information fusion deals with the case where each source provides knowledge by means of a possibility distribution [12], [13]. In this case, it is assumed that the different sources have to cope with imprecision and/or incomplete knowledge and the key question is how uncertainty can be processed when dealing with different sources, that can provide conflicting information.

Acknowledgment: This work is supported by the Flemish Fund for Scientific Research (FWO-Vlaanderen).

In this paper, information fusion is investigated in the context of coreferent objects. An object is hereby axiomatically understood as a piece of data that describes an entity in the real world. The real world is hereby modeled as a reference universe $\mathcal{E}$. Two objects that describe the same entity are called coreferent objects. Detection of such coreferent objects has many important applications, for example in ETL-processes, identification tasks, cleansing of file systems and websites,... In large information systems, the presence of coreferent objects causes inefficient storage and leads to incorrect statistics upon querying, which makes the detection of coreferent objects of utmost importance. It thus comes as no surprise that the detection of coreferent objects has been the topic of many research papers (see [14] for an overview). Interestingly, the problem of *processing* coreferent objects is not as deeply investigated as the detection itself. Merging of non-quantitative objects is not well understood and the properties of good object merging functions are not yet investigated in the context where these objects are coreferent. This paper contributes to filling this gap by investigating a framework for object merging, that allows processing of coreferent objects. Once such a framework for object merging is obtained, it can be used in any of the above mentioned applications of coreference detection.

The remainder of the paper is structured as follows. In Section II, some preliminary notations and definitions are given. In Section III, the properties of merge functions for objects in general, under the premises that the objects are coreferent, are studied. Merge functions for both atomic and complex objects are proposed and the aforementioned properties are evaluated. Finally, in Section IV, the most important contributions of this paper are summarized.

II. Preliminaries

A. Objects

A fundamental concept in this paper is that of an object. An object is axiomatically defined as a piece of data that describes an entity. A distinction is made between atomic and complex objects. Atomic objects are objects of which the universe is non compound, while complex objects belong to a universe O that is composed of non compound universes, i.e. $O = U_1 \times ... \times U_n$. The appropriate universe of entities is denoted

as $\mathcal{E}$ and the link between objects and entities is formalized by a surjective function $\rho : O \to \mathcal{E}$. Objects that refer to the same entity in $\mathcal{E}$ through ρ are said to be coreferent. Formally:

$$\forall(o_1, o_1) \in O^2 : (o_1 \leftrightarrow o_2) \Leftrightarrow (\rho(o_1) = \rho(o_2))$$

The universe of an object is always equipped with a label function $l : O \to \mathcal{L}$, where $\mathcal{L}$ represents the appropriate set of labels. The label of a universe represents the class of entities that objects in the universe are describing. For example, consider $l(\mathbb{R})$ ="temperature", then we know that objects in $\mathbb{R}$ are describing entities of the class "temperature". In addition, complex objects are equipped with a tree-structure in the sense that there exist logical groups of labels that belong together. For example, in objects that describe persons, the universes with label "street", "house number", "postal code" form a logical group, i.e. the address. Formally, for a complex universe O, there exists a function:

$$\lambda : \mathcal{P}\left(\{l\left(U_i\right)\}_{i=1..n}\right) \to \{0, 1\}$$

such that λ indicates for each group of labels, whether these labels form a logical group or not. As the structure that corresponds to λ must be a tree structure, some constraints must be satisfied. The labels themselves must represent leaf nodes and the root node is given by the set of all labels, which means that:

$$\forall i \in \{1, ..., n\} : \lambda(\{l(U_i)\}) = 1$$
$$\lambda(\{l(U_1), ..., l(U_n)\}) = 1$$

Also, the parent child relation must be respected. In terms of λ, this means that for two arbitrary sets of labels, the following constraint must be satisfied:

$$(\lambda(A) = \lambda(B) = 1) \Rightarrow (A \subseteq B \vee B \subseteq A \vee A \cap B = \emptyset)$$

which states that two logical groups A and B are either connected through the ancestor relation or are disjunct.

B. Evaluators

A possibilistic solution for finding coreferent objects consists of finding functions that express uncertainty of coreference by means of *possibilistic truth values* [15], [16], [17], which are possibility distributions over the Boolean domain $\mathbb{B} = \{T, F\}$. Thus, for a given Boolean proposition p, the possibilistic truth value $\tilde{p}$:

$$\tilde{p} = \{(T, \mu_{\tilde{p}}(T)), (F, \mu_{\tilde{p}}(F))\}$$

expresses the possibility that p is true (T) and the possibility that p is false (F). The domain of all possibilistic truth values is denoted $\mathcal{F}(\mathbb{B})$, i.e. the power set of normalized fuzzy sets over $\mathbb{B}$. In what follows, we shall adopt the couple notation for possibilistic truth values, i.e. $\tilde{p} = (\mu_{\tilde{p}}(T), \mu_{\tilde{p}}(F))$. Let us define the order relation $\leq$ on the set $\mathcal{F}(\mathbb{B})$ as follows:

$$\tilde{p} \leq \tilde{q} \Leftrightarrow \begin{cases} \mu_{\tilde{p}}(F) \leq \mu_{\tilde{q}}(F), & \mu_{\tilde{p}}(T) = \mu_{\tilde{q}}(T) = 1 \\ \mu_{\tilde{q}}(T) \leq \mu_{\tilde{p}}(T), & \textbf{else} \end{cases}$$

An evaluator is a function that estimates a possibilistic truth value in order to express uncertainty about coreference [18].

Definition 1: Given a universe of objects O, an evaluator over O is defined as a function E_O:

$$E_O : O^2 \to \mathcal{F}(\mathbb{B})$$

An evaluator compares two objects and yields a possibilistic truth value that expresses both the possibility that the objects are coreferent and the possibility that the objects are not coreferent. An evaluator is *reflexive* if and only if:

$$\forall(o_1, o_2) \in O^2 : (o_1 = o_2) \Rightarrow (E_O(o_1, o_2) = (1, 0))$$

strong reflexive if and only if:

$$\forall(o_1, o_2) \in O^2 : (o_1 = o_2) \Leftrightarrow (E_O(o_1, o_2) = (1, 0))$$

and *commutative* if and only if:

$$\forall(o_1, o_2) \in O^2 : E_O(o_1, o_2) = E_O(o_2, o_1)$$

In what follows, evaluators are always assumed to be commutative and at least reflexive. Finally, an evaluator is called *transitive* if and only if, for every triplet (o_1, o_2, o_3):

$$\text{Nec}(p_{(1,3)} = T) \geq \min\left(\text{Nec}(p_{(1,2)} = T), \text{Nec}(p_{(2,3)} = T)\right)$$
$$\text{Nec}(p_{(1,3)} = F) \geq \min\left(\text{Nec}(p_{(1,2)} = T), \text{Nec}(p_{(2,3)} = F)\right)$$
$$\text{Nec}(p_{(1,3)} = F) \geq \min\left(\text{Nec}(p_{(1,2)} = F), \text{Nec}(p_{(2,3)} = T)\right)$$

with $p_{(i,j)} = E_O(o_i, o_j)$.

C. Multisets

A multiset M derived from a universe U is characterized by a counting function $\omega_M : U \to \mathbb{N}$ ([19]). For $u \in U$, $\omega_M(u)$ represents the number of times that u appears in M. The set of all multisets drawn from a universe U is denoted $\mathcal{M}(U)$. The concept of a subset is extended for multisets as $A \subseteq B \Leftrightarrow \forall u \in U : \omega_A(u) \leq \omega_B(u)$ and the cardinality of a multiset M is given by $|M| = \sum_{u \in U} \omega_M(u)$. Yager defines the following operators for multisets in [19]:

$$\forall u \in U : \omega_{A \cup B}(u) = \max\left(\omega_A(u), \omega_B(u)\right)$$
$$\forall u \in U : \omega_{A \cap B}(u) = \min\left(\omega_A(u), \omega_B(u)\right)$$
$$\forall u \in U : \omega_{A \oplus B}(u) = \omega_A(u) + \omega_B(u)$$

The $\in$-relation applies for multisets as follows: $u \in M \Leftrightarrow \omega_M(u) > 0$. The k-cut of a multiset M is a regular set $M_k = \{u | u \in U \wedge \omega_M(u) \geq k\}$.

III. Object merging

A. Definition and properties

In this Section, the concept of a merge function for arbitrary objects is given and relevant properties in the context of coreferent objects are proposed. A merge function for objects is defined as follows.

Definition 2 (Merge function): For a universe O, a merge function over O is a function:

$$\varpi_O : \mathcal{M}(O) \to O$$

A merge function thus takes a multiset of objects and produces a single object as a result. It can be seen that this very general definition allows the construction of many irrelevant merge functions. Therefor, the properties that make up a good merge function are investigated in the following. The first property that is proposed here, is *idempotency*.

Property 1 (Idempotency): A merge function ϖ_O is idempotent if and only if:

$$\forall o \in O : \varpi_O(\{o, o, ..., o\}) = o$$

The idempotency property reflects that, in case of total agreement, the resulting object should be the one that all sources agree upon. Idempotency is a very natural property and should always be satisfied. The second property, *monotonicity*, is borrowed from the axioms of aggregation operators.

Property 2 (Monotonicity): Consider two arbitrary multisets M_1 and M_2 drawn from a universe O. A merge function ϖ_O is monotone if and only if, for any one-to-one mapping f between M_1 and M_2 such that:

$$\forall (o_1, o_2) \in f : o_1 \leq o_2$$

it holds that:

$$\varpi_O(M_1) \leq \varpi_O(M_2)$$

Monotonicity, although quite natural and useful, is not easy to satisfy in the general case. It is for example not always clear which order function $\leq$ should be used to compare objects. As an example, consider the case of strings. The natural order of strings is the alphabetical order. However, next to this total order, there exist other partial orders, such as the orders induced by the substring relation and the subsequence relation. Considering the fact that the alphabetical order and the substring order are not equivalent, it can be questioned which of these orders should be taken into account. Also, the construction of an order relation for complex objects is not always clear. A third property is called *preservation*.

Property 3 (Preservation): Given a universe O and a merge function ϖ_O, then ϖ_O is preservative if and only if:

$$\forall M \in \mathcal{M}(O) : \varpi_O(M) \in M$$

In words, a preservative merge function chooses one of the objects. Preservation of a merge function can be an interesting property in the context of coreferent objects for several reasons. Firstly, preservation ensures *traceability*, which means that a merged object can be traced back to the original source(s). In many practical environments, such traceability is of crucial importance. Secondly, when dealing with complex objects, an arbitrary mix of sub-objects might not always make sense. For example, when merging two different addresses, it does not make sense to choose the street of the first address and the house number of the second address, because there is no verification at all whether the resulting address exists. In this case, non-preservative merging can result in an object that does not at all refers to the entity that was referred to by the set of input objects. In order to better explain the context in which preservation is a useful property, consider the object

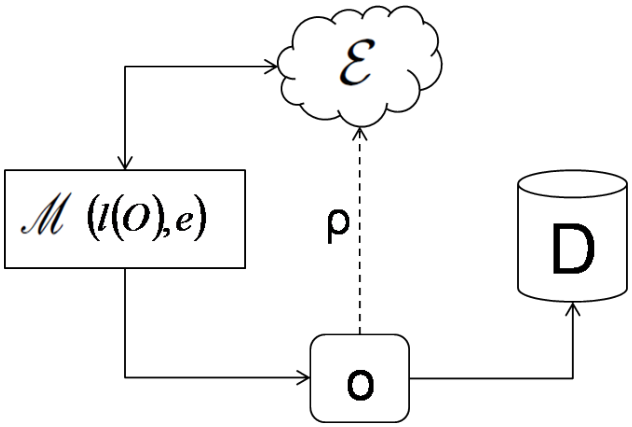

Fig. 1. The object creation process

creation process, as illustrated in Figure 1. This Figure shows that an object o is created when it needs to be inserted into a data source D. Therefor, an entity e is *measured* by a process $\mathscr{A}$. In the general case of complex objects, measurement of an entity implies measurement of several *attributes* of the entity, where each attribute is denoted by the label of a sub-universe U_i. The relevance of $\mathscr{A}$ with respect to merging, is that we can distinguish between the case where $\mathscr{A}$ measures precisely and the case where $\mathscr{A}$ does not measure precisely. More specific, whenever $\mathscr{A}$ can precisely measure objects, then coreferent objects that are *not equal* differ because of a reason other than imprecision. As such, it can be argued that any inconsistency upon merging time is due to heterogeneity in the representation of the entity that was measured. Thus, as the measurements at hand are known to be precise measures, it is acceptable to choose one of the objects, i.e. by choosing the one that is known as the *most common* representation. Hence, when dealing with precise measurements, preservation is a desired property. In the other case, where $\mathscr{A}$ is not able to measure entities with full precision, none of the coreferent objects under consideration might be a common description of the entity. To better understand the difference between both situations, consider the following example. Suppose a company manages a website with a geographical information system (GIS) where website visitors can search for interesting locations, i.e. points of interest, within a certain region. A problem hereby is the discovery of interesting locations. Therefor, the company adopts a community-driven model, where website visitors can enter their own points of interest. To do so, they need to pinpoint a location on a map, which is stored in two attributes called latitude and longitude. Next, they can attach a name to this location. In this setting, it is clear that the measurement of attributes latitude and longitude suffers from imprecision for several reasons. It is perhaps not possible to describe the location in one point because it is an area or the exact location can be unknown to the users. Yet in another scenario, the physical properties of the input device can put a constraint on the precision of measurement. However, measurement of the name attribute can be done

precisely. A fourth property, that is mentioned in many works dealing with merging of propositional belief bases, is called the *majority rule* [8], [10].

Property 4 (Majority rule): Let O be a universe and let ϖ_O be a merge function, then ϖ_O satisfies the majority rule if and only if:

$$\forall M \in \mathcal{M}(O) : \exists o \in O : \omega_M(o) > \frac{|M|}{2} \Rightarrow (\varpi_O(M) = o)$$

The majority rule is an interesting property in the context of coreference as it implies that a majority of sources postulates that a certain object is the correct representation of the entity. However, the majority of sources might not always know the most common representation of an entity. Therefor, a contradictory but interesting property is called *majority independence*. To formulate this property, we first define the suppression operator for multisets.

Definition 3: Let M be a multiset drawn from the universe O. The k-suppression of M is a multiset $\langle M \rangle_k$ such that

$$\omega_{\langle M \rangle_k}(o) = \begin{cases} 1 & , 0 < \omega_M(o) < k \\ \omega_M(o) & , \textbf{else} \end{cases}$$

Property 5 (Majority independence): Given a universe O and a merge function ϖ_O, then ϖ_O is k-majority independent if and only if:

$$\forall M \in \mathcal{M}(O) : \varpi_O(M) = \varpi_O(\langle M \rangle_k)$$

B. Merging of atomic objects

We now focus on merge functions ϖ_U where U is an atomic universe. Recall that the context in which ϖ_U is to be used, is that of coreference. As such, we can assume that upon merging time, an evaluator E_U is available. Let M be a multiset of coreferent objects that are identified by a coreference detection framework. Then, for each object $u \in M$, $|M|$ possibilistic truth values can be calculated by comparing u with all objects in M. Due to reflexivity of E_U, the possibilistic truth value $(1, 0)$ occurs at least $\omega_M(u)$ times. As such, a collection of possibilistic truth values is obtained where each $\tilde{p}$ indicates the uncertainty about the proposition that two objects are coreferent. In [20], a method is proposed to construct a possibility distribution $\pi_{\mathbb{N}}$ (a fuzzy integer) from a collection of possibilistic truth values $\tilde{P}$. Hereby, $\pi_{\mathbb{N}}(k)$ indicates the possibility that k propositions in P are true. The method explained in [20] is the following.

Definition 4: Let P be a set of independent Boolean propositions and let $\tilde{P}$ be the set of corresponding possibilistic truth values. The quantity of true propositions in P is given by the possibility distribution $\pi_{\mathbb{N}}$ such that:

$$\pi_{\mathbb{N}}(k) = \min\big(\sup\{\alpha \in [0,1] || \{p \in P | \mu_{\tilde{p}}(T) \geq \alpha\}| \geq k\}, \sup\{\alpha \in [0,1] || \{p \in P | \mu_{\tilde{p}}(F) < \alpha\}| \geq k\}\big)$$

Definition 4 states that the possibility $\pi_{\mathbb{N}}(k)$ is the minimum of the possibility that at least k propositions are true and the possibility that at most $|P| - k$ propositions are false. From this point of view, $\pi_{\mathbb{N}}(k)$ can be calculated by adopting the following notations. For a set $\tilde{P}$, let $\tilde{p}_{(i)^T}$ denote the i^{th} largest possibilistic truth value with respect to the order relation defined in Section II. The following then holds:

$$\pi_{\mathbb{N}}(k) = \begin{cases} \mu_{\tilde{p}_{(k)^T}}(F) & , k = 0 \\ \mu_{\tilde{p}_{(k)^T}}(T) & , k = |M| \\ \min\left(\mu_{\tilde{p}_{(k)^T}}(T), \mu_{\tilde{p}_{(k+1)^T}}(F)\right) & , \textbf{else} \end{cases}$$

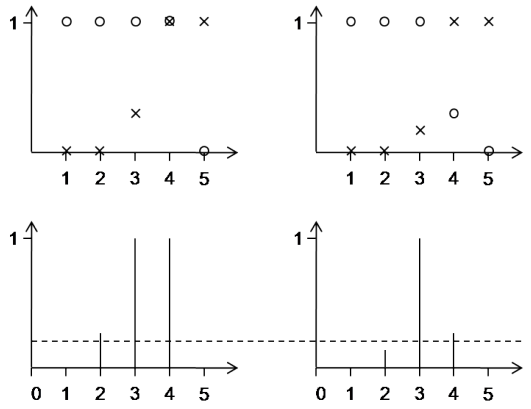

Fig. 2. Fuzzy integers derived from possibilistic truth values

Figure 2 shows two example sets of possibilistic truth values, where $\circ$ denotes the possibility of T and $\times$ denotes the possibility of F. The derived possibility distributions $\pi_{\mathbb{N}}$ are shown below the possibilistic truth values. Note that the membership functions of the derived fuzzy integers $\pi_{\mathbb{N}}$ are always *convex*.

Applying this method allows us to express the number of coreferent objects, *according to* the evaluator E_U. Hence, although we already know that objects in M are coreferent, the distributions $\pi_{\mathbb{N}}$ express the uncertainty about this statement, at least, according to the evaluator E_U. In light of this, the result of $\varpi_U(M)$ should be the object which has the highest number of coreferent objects according to E_U. We then obtain a method where the uncertainty model of E_U is used to choose the best representative. For this purpose, a method for comparison of fuzzy integers is required. Many methods have been proposed. The most common technique is to *defuzzify* the fuzzy integer, for example by means of the center of gravity [1]. Fuzzy integers are then compared by comparing the results of defuzzification. The method that we shall adopt here, is not based on defuzzification, but is rather possibilistic in nature. We propose two order relations for fuzzy integers, one constructed from the viewpoint of possibility and one constructed from the viewpoint of necessity.

*Definition 5 (*sup*-order of fuzzy integers):* For two fuzzy integers, $\tilde{n}$ and $\tilde{m}$, the order relation $\prec_{\sup}$ is defined as:

$$\tilde{n} \prec_{sup} \tilde{m} \Leftrightarrow \sup \tilde{n}_\alpha < \sup \tilde{m}_\alpha$$

Hereby, $\tilde{n}_\alpha$ is the α-cut of $\tilde{n}$ where α is chosen such that:

$$\alpha = \sup\{x | \sup \tilde{n}_x \neq \sup \tilde{m}_x\}$$

Definition 6 ($\inf$*-order of fuzzy integers):* For two fuzzy integers, $\tilde{n}$ and $\tilde{m}$, the order relation $\prec_{\inf}$ is defined as:

$$\tilde{n} \prec_{\inf} \tilde{m} \Leftrightarrow \inf \tilde{n}_\alpha < \inf \tilde{m}_\alpha$$

Hereby, $\tilde{n}_\alpha$ is the α-cut of $\tilde{n}$ where α is chosen such that:

$$\alpha = \sup\{x | \inf \tilde{n}_x \neq \inf \tilde{m}_x\}$$

The $\sup$-order of fuzzy integers searches for the highest α, such that the α-cuts have a different supremum and then chooses the fuzzy number for which the α-cut has the higher supremum. It can be seen that this method is equivalent to first searching the fuzzy integers that have the maximal k, say $k_{\max}$, for which $\pi_{\mathbb{N}}(k_{\max}) = 1$. If multiple fuzzy integers exist, the decision is taken by leximax of the sequence $\pi_{\mathbb{N}}(k_{\max}+1), ..., \pi_{\mathbb{N}}(|M|)$. The dual is true for $\prec_{\inf}$. Note that both $\prec_{\sup}$ and $\prec_{\inf}$ are partial orders. If multiple fuzzy numbers are equivalent, a random choice is made. Note that two non-equal convex fuzzy integers are always comparable by either $\prec_{\inf}$ or $\prec_{sup}$. Consider the fuzzy integers shown in Figure 2. The order relation $\prec_{\sup}$ denotes the leftmost fuzzy integer as the largest, because 1-cut of the leftmost fuzzy integer has a higher supremum (4) than the rightmost (3). However, the order relation $\prec_{\inf}$ denotes the rightmost fuzzy integer as the largest, because the 0.2-cut (denoted by the dashed line) of the leftmost fuzzy number has a lower infimum than the 0.2 cut of the rightmost fuzzy integer.

Based on the order of fuzzy integers, it is possible to define a merge function driven by an evaluator for atomic universes, denoted ϖ_U^k.

Definition 7 (Evaluator driven ϖ_U^k): Let U be a atomic universe and E_U an evaluator over U. A merge function driven by the evaluator E_U of order k is a merge function ϖ_U^k such that:

$$\varpi_U^k(M) = \arg\max_{u \in M} \pi_{\mathbb{N}}^u$$

where $\pi_{\mathbb{N}}^u$ is a possibility distribution obtained from the multiset of possibilistic truth values $\tilde{P}_u$ where:

$$\forall u' \in M : \omega_{\tilde{P}_u}(E_U(u,u')) = \omega_{\langle M \rangle_k}(u')$$

It can now be verified easily that ϖ_U^k, is k-majority independent. Indeed, ϖ_U^k suppresses multiplicity below k. In what follows, we shall study the most interesting case where $k = 1$. Therefor, for simplicity of notation, we will drop the superscript $(.)^k$ in the sequel. Considering the other properties of merge functions, evaluator driven merge functions are preservative by definition, and thus also idempotent. The majority rule is not easily satisfied. It can be shown with simple counter examples that, in the general case, ϖ_U does not satisfy the majority rule, independent of the ordering relation for fuzzy integers used. However, there exist some cases in which the majority rule is in fact satisfied.

Theorem 1: A merge function ϖ_U, driven by a strong reflexive and transitive evaluator E_U satisfies the majority rule if it orders fuzzy integers with $\prec_{\inf}$.

Proof: Assume a the element of M, that has the majority. This means that $\omega_M(a) > \left\lfloor \frac{|M|}{2} \right\rfloor$. Notice that:

$$\forall u \in M : |\{\tilde{p} \in \tilde{P}_u | \mu_{\tilde{p}}(F) \neq 1\}| = \arg\min \pi_{\mathbb{N}}^u(k) = 1$$

which means that the infimum of the 1-cut of $\pi_{\mathbb{N}}(k) = 1$ is equal to the number of possibilistic truth values with possibility for F lower than 1. Thus, if:

$$\forall u \in M \backslash \{a\} : \mu_{E_U(u,a)}(T) < 1$$

then $\pi_{\mathbb{N}}^a$ has the strictly largest infimum of the 1-cut. In the other case where there exists a multiset $C \subset M$ such that:

$$\forall u \in C : \mu_{E_U(u,a)}(T) = 1$$

then we see that $\forall u \notin C : \pi_{\mathbb{N}}^u \prec_{\inf} \pi_{\mathbb{N}}^a$ due to the previous case. For elements in C we have that:

$$\forall u \in C : \pi_{\mathbb{N}}^u(|C|) = 1$$

We thus have to show that for each $u \in C$, $\pi_{\mathbb{N}}^u$ dominates $\pi_{\mathbb{N}}^a$ in the index set $\{1, 2, ..., |C|\}$. This follows from the fact that, on the one hand for every b and c in C, different from a, we have that:

$$E_U(b,c) < E_U(a,c) \Rightarrow E_U(a,b) = E_U(a,c)$$

due to transitivity and on the other hand that:

$$\forall k \in \{1, ..., \lfloor |M|/2 \rfloor\} : \pi_{\mathbb{N}}^a(k) = 0$$

■

C. Merging of complex objects

In this Section, merge functions for complex objects $o \in O$ are investigated. A possible strategy in doing so is to consider an evaluator E_O and to construct merge functions for complex universes as explained in the previous Section. A different way of defining merge functions for complex objects is to combine the *projection* operator on the compound universe O with merge functions for the atomic universes. Doing so, yields the following definition of a composite merge function.

Definition 8 (Composite merge function): Consider a complex universe $O = U_1 \times ... \times U_n$. A composite merge function ϖ_O over O is defined as:

$$\varpi_O : \mathcal{M}(O) \to O$$

where:

$$\varpi_O(M) = \left(\varpi_{U_1}\left(\text{Proj}_1\left(M\right)\right), ..., \varpi_{U_n}\left(\text{Proj}_n\left(M\right)\right)\right)$$

and where $\text{Proj}_i(M) \in \mathcal{M}(U_i)$ such that:

$$\omega_{\text{Proj}_i(M)}(u) = \sum_{o \in M \wedge o_i = u} \omega_M(o)$$

In case of a composite merge function it is now investigated which properties, satisfied by the functions ϖ_{U_i}, are inherited by ϖ_O. It can be verified that if ϖ_O is composed of idempotent functions ϖ_{U_i}, then ϖ_O is idempotent. If the ϖ_{U_i} satisfy the majority rule, then so does ϖ_O and if the functions ϖ_{U_i} are k-majority independent, then so is ϖ_O. A property that requires

more attention, is preservation. Indeed, if all ϖ_{U_i} are preservative, then a composite ϖ_O is not bound to be preservative. However, due to preservation of the merge functions ϖ_{U_i}, the resulting object $\varpi_O(M)$ can be projected onto each sub-universe. As such, we can count the number of times each input object contributes to the resulting object and we can find the input object that contributes the most. Formally, it is possible to define a merge function $\varpi_O^* : \mathcal{M}(O) \rightarrow O$ such that:

$$\varpi_O^*(M) = \arg\max_{o \in M} (\omega_M(o) \cdot |\{i | o_i = \varpi_O(M)_i\}|)$$

where ϖ_O is a merge function. Then, ϖ_O^* is a preservative merge function induced by a composite merge function. The advantage of preservation for complex objects is already pointed out in the previous Section. However, preservation in the case of complex objects can have the disadvantage that it is not robust. Indeed, if the input object that is chosen contains an error in one sub-universe, but has preferable values in the other universes, preservation implies that this error is not corrected. Therefor, in the context of complex objects, a more interesting property is λ-preservation.

Property 6 (λ-preservation): Assume a complex universe $O = U_1 \times ... \times U_n$ and a merge function ϖ_O. Then ϖ_O is λ-preservative with respect to a λ-partition $\{P_j\}_{j=1..k}$ of labels, if it is preservative with respect to each P_j. Hereby, a λ-partition satisfies:

$$\bigcup_{j=1}^{k} P_j = \{l(U_i)\}_{i=1..n}$$
$$\forall j \in \{1, ..., k\} : \lambda(P_j) = 1$$

In an extreme case, each sub-universe is preserved independently from the others, i.e. in each sub-universe, the merged object has a value that belongs to one of the input objects, but for each sub-universe, this object can differ. As such, with respect to this trivial partition, a composite merge ϖ_O composed from preservative merge functions ϖ_{U_i} is λ-preservative. The strength of the property depends on the λ-partition for which preservation holds. When λ-preservation is obtained with respect to logical, non singleton groups of sub-universes, the risk of constructing an object of low quality is lowered. On the other hand, by not selecting an object of the input set, the merge function has the ability of correcting errors in the input set.

IV. Conclusion

Information fusion is a research area with many applications, such as heterogeneous databases, multi-agent systems and group decision making. In this paper, information fusion is investigated in the context of coreferent objects. We start with defining merge functions in general. Next, interesting properties in the context of coreferent objects are given. We then define merge functions for atomic objects and we show which properties are satisfied by these merge functions. Next, it is shown how composite merge functions can be defined by means of atomic merge functions and it is shown how properties of the atomic merge functions can be satisfied for the composite merge functions. The proposed framework of merge functions contributes to automated processing of coreferent objects.

References

[1] Didier Dubois and Henri Prade, *Fundamentals of Fuzzy Sets*. Kluwer Academic, 2000.

[2] Jozo Dujmovic, "A generalization of some function in continuous mathematical logic - evaluation function and its applications," in *Proceedings of the Informatica Conference*, Yugoslavia, 1973.

[3] Ronald Yager, "On ordered weighted averaging aggregation operators in multicriteria decision making," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 18, no. 1, pp. 183–190, 1988.

[4] Ronald Yager and Alexander Rybalov, "Uninorm aggregation operators," *Fuzzy Sets and Systems*, vol. 80, no. 1, pp. 111–120, 1996.

[5] A Borgida and T Imielinski, "Decision making in committees: a framework for dealing with inconsistency and non-monotonicity," in *Proceedings Workshop of Nonmonotonic reasoning*, 1984, pp. 21–32.

[6] Chitta Baral, Sarit Kraus, and Jack Minker, "Combining multiple knowledge bases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 3, no. 2, pp. 208–220, 1991.

[7] Chitta Baral, Sarit Kraus, Jack Minker, and V Subrahmanian, "Combining knowledge bases consisting of first-order theories," *Computational Intelligence*, vol. 8, no. 1, pp. 45–71, 1992.

[8] Jinxin Lin and Alberto Mendelzon, "Knowledge base merging by majority," in *In Dynamic Worlds: From the Frame Problem to Knowledge Management*. Kluwer, 1994.

[9] ——, "Merging databases under constraints," *International Journal of Cooperative Information Systems*, vol. 7, no. 1, pp. 55–76, 1998.

[10] S Konieczny and R Pérez, "Merging information under constraints: a logical framework," *Journal of Logic and Computation*, vol. 12, no. 1, pp. 111–120, 2002.

[11] M Bright, A Hurson, and S Pakzad, "A taxonomy and current issues in multidatabase systems," *Computer*, vol. 25, no. 3, pp. 50–59, 1992.

[12] S Sandri, Didier Dubois, and H Kalfsbeek, "Elicitation, assessment and pooling of expert judgements using possibility theory," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 3, pp. 313–335, 1995.

[13] Sebastien Destercke, Didier Dubois, and Eric Chojnacki, "Possibilistic information fusion using maximal coherent subsets," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 1, pp. 79–92, 2009.

[14] Ahmed Elmagarmid, Panagiotis Ipeirotis, and Vassilios Verykios, "Duplicate record detection: A survey," *IEEE Transactions on Knowledge and Data engineering*, vol. 19, no. 1, pp. 1–16, 2007.

[15] Henri Prade, "Possibility sets, fuzzy sets and their relation to lukasiewicz logic," in *Proceedings of the International Symposium on Multiple-Valued Logic*, 1982, pp. 223–227.

[16] Antwan Van Schooten, "Ontwerp en implementatie van een model voor de representatie en manipulatie van onzekerheid en imprecisie in databanken en expert systemen," Ph.D. dissertation, Ghent University, 1988.

[17] Gert De Cooman, "Evaluatieverzamelingen en - afbeeldingen. een ordetheoretische benadering van vaagheid en onzekerheid," Ph.D. dissertation, Faculteit Toegepaste Wetenschappen, Universiteit Gent, 1993.

[18] Antoon Bronselaer, Axel Hallez, and Guy De Tré, "A possibilistic view on set and multiset comparison," *Control and Cybernetics*, vol. 38, no. 2, pp. 341–366, 2009.

[19] Ronald Yager, "On the theory of bags," *International Journal of General Systems*, vol. 13, no. 1, pp. 23–27, 1986.

[20] Axel Hallez, Guy De Tré, Jörg Verstraete, and Tom Matthé, "Application of fuzzy quantifiers on possibilistic truth values," in *Proceedings of the Eurofuse Workshop on Data and Knowledge Engineering*, Warshau, Poland, 2004, pp. 252–254.