Control and Cybernetics

vol. **38** (2009) No. 2

A possibilistic view on set and multiset comparison*

by

Antoon Bronselaer, Axel Hallez and Guy De Tré

Dept. of Telecommunications and Information Processing,
Ghent University
Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium

Abstract: Comparative evaluation operators for sets and multisets are proposed from a possibilistic point of view. In general, an evaluator estimates the possibility of (non) co-reference of two arbitrary (sub)-objects. Such operators can be used in a hierarchical possibilistic framework for finding co-referent objects with a complex structure. This paper first discusses properties of evaluators in general and continues with studying operators for sets and multisets, thereby making a clear distinction between hard and soft evaluators. Hard evaluators are based on evaluation of derived (multi)sets, while soft evaluators use a low level evaluator to incorporate co-reference at element level. The two important parts of such a soft evaluator are an injective element mapping and an aggregation function. An algorithm to provide the injective mapping is presented and discussed. For the aggregation step, ordered weighted conjunction is studied by introducing parameterized fuzzy quantifiers to calculate weight vectors. An advanced learning strategy is introduced to train the optimal parameter matrix.

Keywords: evaluator, co-reference, multisets, sets.

1. Introduction

In everyday life, storage of information has become of key importance. Whether using a high-tech database or a simple paper sheet, each process of data storage is based on the principle of *description of real world phenomena*. Since the introduction of databases, several means of representing and storing data in a *structural* way have been proposed (relational databases, XML, OO-environment,...). In the scope of this paper, such a structural description of a real world entity is called an *object*. Duplicate objects are two objects that represent or describe the same real world entity. For that reason, they are called co-referent objects and the problem of finding them is called the *co-reference problem* (the term co-reference is introduced in Cohen, 1998). Detection of co-referent objects has been investigated extensively in the past decades. In applications of data(base)

*Submitted: March 2008; Accepted: July 2008.

merging, it is of vital importance to avoid duplicate storage and inconsistencies, as they both lead to inefficient data management and could introduce ambiguity in the data.

This paper deals with objects that have a predefined structure. A probabilistic model to deal with this problem was given by Fellegi and Sunter (1969). In more recent work, a hierarchical framework for object matching was introduced (Hallez and De Tré, 2007), which is a generic framework in the sense that the domain in which results are expressed, is left unspecified. The general expression domain is due to the fact that object matching is a more general problem than the co-reference problem. For example, comparing a query object with a stored object is a matter of preference expression, while finding two co-referent objects is an uncertain boolean problem, as will be pointed out in the following. Nevertheless, in both cases two objects are compared and a result is given, expressing the answer to the question posed. The mentioned framework is called hierarchical as it exploits a hierarchy of operators to infer the final result. Equipping the hierarchical framework with the domain of possibilistic truth values, leads to a possibilistic model for the co-reference problem.

The choice for a possibilistic approach on co-reference is justified by several reasons.

First of all, objects are (real world) entity descriptions and co-referent objects describe *equal* entities, so co-reference is a boolean matter. Either two objects are co-referent, or not, and there is no such thing as a co-reference degree. However, due to imperfections in the data, entities can be described in different ways, which implies that it can be uncertain whether objects are co-referent or not. An elegant tool to express such uncertainty about boolean propositions are possibilistic truth values (Prade, 1982), which are epistemological values that describe knowledge or belief about truth values (i.e. they are not truth values themselves). Thus, the possibilistic approach presented here expresses the possibility (i.e. the belief) that two objects describe the same entity or not. As similarity relations are not always compatible with an intuitive assignment of such possibilities, the presented approach has a benefit over similarity relations.

Secondly, possibilistic truth values have been recognized in the past as the desired machinery to deal with linguistic uncertainty (De Cooman, 1995), which means that the possibilistic approach can deal with objects containing linguistic terms.

Thirdly, missing data (unknown values or non-existing values) imply uncertainty in the reasoning process. Using possibilistic truth values allows for elegant reasoning about missing data.

The possibilistic model for object matching was elaborated in Bronselaer and De Tré (2007, 2008), where the focus was mainly on aggregation of (intermediate) results and preference modeling. However, in order to support reasoning, the possibilistic model requires also possibilistic evaluators for low-level comparison, i.e. attribute comparison, to deliver initial possibilistic truth values. These operators express the uncertainty about the boolean truth value of the proposition that two attribute values refer to the same real world value. As these operators deliver the input for the reasoning process and thus determine the outcome of this reasoning to a large extent, they are of utmost importance. Still, their development is, up till now, not deeply investigated. Therefor, this work offers a prototype for such evaluation operators in the possibilistic framework. Any domain can be equipped with such an evaluator, but especially important cases are: numerical data, strings, (multi)sets and linguistic terms (modeled by possibility distributions). The semantics of the operators are discussed and the usefulness in practice of some properties is given. Next, evaluators are constructed for the case of sets and multisets. The choice of elaborating on these two datatypes is justified by noting that collection datatypes have interesting applications. Firstly, in databases and object oriented languages, many-valued attributes often occur, often modeled as (multi)sets or lists. Examples of such attributes are 'spoken languages', 'hobbies', 'friends',... Secondly, detection of co-referent multisets has a very interesting application in the detection of co-referent strings, where a tokenization function splits a string into a multiset of substrings. An explicit difference between *hard* and *soft* evaluators for (multi)sets is introduced. Hard evaluators use derived (multi)sets (such as intersection and union) to formulate a result, which implies that on element level, strict equality is used. Soft evaluators are a generalization of hard evaluators in the sense that they assume that non-equal elements can be co-referent. They use an additional evaluator on the element level, that generates a sequence of intermediate results and an aggregation function to infer a final result. The use of ordered weighted conjunction (OWC) for this latter purpose is investigated by introducing parameterized fuzzy quantifiers specifically designed for comparison of (multi)sets. These quantifiers can be used to calculate the weight vector of the OWC.

The paper is structured as follows. In Section 2, some basic concepts are introduced, followed by a brief summary of previous work related to this paper in Section 3. Section 4 describes a general possibilistic evaluator, which leads to an evaluator for sets in Section 5 and multisets in Section 6. Finally, a guideline for future work is given in Section 7 and the main contributions of this work are summarized in Section 8.

2. Preliminaries

2.1. Possibilistic truth values

A possibilistic truth value (PTV) is a possibility distribution, represented by a fuzzy set, defined over the set of boolean values $I = \{T, F\}$, where T represents true and F represents false (Zadeh, 1978; Prade, 1982). PTVs are used to express the uncertainty about the boolean value of a proposition. In contradiction to what their name might imply, PTVs are not truth values, but epistemological values. This means that they describe a state of knowledge or belief about the truth value of a proposition. Let P denote the set of all propositions, then each $p \in P$ can be associated with a PTV $\tilde{p} = \{(T, \mu_{\tilde{p}}(T)), (F, \mu_{\tilde{p}}(F))\}$, where $\mu_{\tilde{p}}(T)$ represents the possibility that p is true and $\mu_{\tilde{p}}(F)$ represents the possibility that p is false. The set of all PTVs is denoted $\tilde{\wp}(I)$. It is assumed that $\max(\mu_{\tilde{p}}(T), \mu_{\tilde{p}}(F)) = 1$, which reflects our assumption that the universe I is large enough to express the truth value of p. Within the framework of PTVs, it is possible to define generalizations $\tilde{R}$ of order relations R as follows:

$$\tilde{p}_1 \, \tilde{R} \, \tilde{p}_2 \Leftrightarrow \begin{cases} \mu_{\tilde{p}_2}(F) \; R \; \mu_{\tilde{p}_1}(F), & \text{if} \quad \mu_{\tilde{p}_2}(T) = \mu_{\tilde{p}_1}(T) = 1 \\ \mu_{\tilde{p}_1}(T) \; R \; \mu_{\tilde{p}_2}(T), & \text{otherwise}. \end{cases}$$

The framework of PTVs also provides generalizations of boolean operators in order to aggregate uncertainty about boolean values. Within the scope of this paper, the most important operator is a generalization of *conjunction*:

$$\begin{aligned} &\tilde{\wedge} : \tilde{\wp}(I)^2 \to \tilde{\wp}(I) : \\ &\qquad \tilde{p}\tilde{\wedge}\tilde{q} \mapsto \{(T, t(\mu_{\tilde{p}}(T), \mu_{\tilde{q}}(T))), (F, s(\mu_{\tilde{p}}(F), \mu_{\tilde{q}}(F)))\} \end{aligned}$$

where (t, s) is a t-norm/t-conorm pair, such that the possibilistic variables $\tilde{p}$ and $\tilde{q}$ are t-independent (De Cooman, 1995). In what follows, we will make use of the couple notation for PTVs, which represents a PTV $\tilde{p}$ as $(\mu_{\tilde{p}}(T), \mu_{\tilde{p}}(F))$.

2.2. Multisets

Part of this work will focus on the comparison of multisets, which are an extension of regular sets. In the remainder of this work, a multiset M derived from a universe U is characterized by a counting function $C_M : U \to \mathbb{N}$ (Yager, 1986). For $u \in U$, $C_M(u)$ represents the number of times u appears in M. The set of all multisets drawn from a universe U is denoted $\mathcal{M}(U)$. Yager (1986) defines some extensions of set operators for multisets:

$$\begin{aligned} &\forall u \in U : C_{A \cup B}(u) = \max(C_A(u), C_B(u)) \\ &\forall u \in U : C_{A \cap B}(u) = \min(C_A(u), C_B(u)) \\ &\forall u \in U : C_{A \ominus B}(u) = \max(C_A(u) - C_B(u), 0) \\ &\forall u \in U : C_{A \oplus B}(u) = C_A(u) + C_B(u). \end{aligned}$$

Next, for these operators the concept of subset is extended for multisets as $A \subset B \Leftrightarrow \forall u \in U : C_A(u) \leq C_B(u)$ and the cardinality of a multiset M can be computed by $\forall M \in \mathcal{M}(U) : |M| = \sum_{u \in U} C_M(u)$.

3. Related work

As mentioned in the introduction, Fellegi and Sunter (1969) were the first to give a formal solution for the duplicate detection problem in databases. Their solution is based on probability theory and assumes records consisting of n

fields. When comparing two records r_1 and r_2 the field values of both records are first compared to each other, resulting in a vector $\underline{x} = [x_1, ..., x_n]$ of attribute comparisons. Hereby, x_i represents the [0,1]-valued similarity of the values of the i^{th} field. Next, a Bayesian network outputs a decision on whether the two records are duplicates, based on $\underline{x}$.

Several methods are proposed to estimate the conditional probabilities of the Bayesian network. Jaro (1989) assumed conditional independence between the probabilities to be estimated and suggested the use of an expectation maximization (EM) algorithm. Winkler (1993) generalized this idea to the case where the conditional independence assumption is violated. Du Bois (1969) pointed out the importance of dealing with missing data.

An alternative to Bayesian modeling is a rule based approach, which has been studied extensively in Wang and Madnick (1989), Hernandez and Stolfo (1998), Tejada, Knoblock and Minton (2001), and Koyuncu and Yazici (2001). While the previous works focus on database records, other works deal with more complex objects, such as XML-documents and OO-environments (see Marin et al., 2003; Doan et al., 2003; and Weis and Naumann, 2004).

In more recent work, Hallez and De Tré (2007), a new model was proposed to deal with a more general problem called *object matching*, where objects are assumed to be *structured*. Taking the structure of objects into account during comparison results in a more natural reasoning process, which is why this model is adopted here. Next, to the work mentioned here, a large body of literature on co-referent objects exists and it is not feasible to describe every paper here. Two good overview papers are Winkler (2006) and Elmagarmid, Ipeirotis and Verykios (2007).

A significant part of this paper focuses on comparison of sets and multisets. The first work concerning this topic is due to Jaccard (1908), who introduced the well known Jaccard index for sets. This index was generalized by Tversky in (1977). Dubois and Prade (1982) introduced comparison indexes in a fuzzy set theoretic framework. In Matthé et al. (2006) an algorithm is provided that takes similarities between elements into account. However, the outcome of this algorithm is not unique. In addition, Matthé et al. (2006) provides an extension of the Jaccard index, whereas this paper shows how ordered weighted conjunction can be used in order to provide a final result. For a complete overview concerning comparison indexes for (fuzzy) sets and their properties, the reader is referred to Cross and Sudkamp (2002).

4. Possibilistic evaluation

As mentioned before, this paper contributes to a possibilistic approach on the coreference problem. Therefor, we begin by describing this model and its purpose.

The term ‘object’ refers to an arbitrarily complex description of a structured entity. Examples of such entities are cars, persons, ... The problem faced in this paper, is the process of finding those pairs of objects that describe the same entity. Such pairs of objects are called *co-referent* objects. It is assumed here that the objects reflect the natural structure of the entities in a hierarchical way, more specific by using a *tree structure*. The scope of this paper concerns objects that share such a predefined structure. A possibilistic solution for this problem is provided next. Given two objects, we have the following affirmative proposition $p =$ "o_1 and o_2 are co-referent", which evaluates to a boolean value. As entity description allows heterogeneous representations of the same entity, non-equal objects can refer to the same entity. Hence, there is an implicit uncertainty about the boolean value of p, which can be modeled by a possibilistic truth value (Section 2). Consequently, finding co-referent objects requires providing the PTV associated with proposition p, which is equivalent to calculating the membership grades of this PTV. These membership grades are computed by comparing the sub-objects defined in the object structure shared by both objects. The basic sub-objects are called the *attributes* and comparing the values of n attributes results in n basic propositions $p_i =$ "o_1 and o_2 have co-referent values for the i^{th} attribute". Attributes are sometimes assumed to be atomic, but this assumption is omitted here in order to support many-valued attributes. The operators that formulate possibilistic statements about such propositions are called *possibilistic comparative evaluation operators* or *evaluators* for short. These statements are combined by using *aggregation operators* for PTVs, which are an extension of their corresponding logical boolean operators.

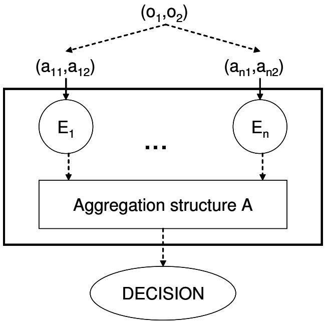

Figure 1. General structure of a comparison scheme

To clarify the possibilistic framework, the general structure of a hierarchical possibilistic comparison scheme is shown in Fig. 1, where A represents an aggregation structure (i.e. a complex tree-structure of aggregation operators) and E_i denotes a possibilistic evaluation operator for the i^{th} attribute.

Little work has been done concerning the development of such evaluators E that directly estimate the possibilities of the boolean value of p_i (in De Cooman, 1995, an example can be found that deals with linguistic terms). For that purpose, we first define a generic form of such an operator.

Definition 1 (Evaluator) *Assume a universe U. For each couple of values $(u, u') \in U^2$, assume an affirmative proposition p = "u and u' are co-referent". The uncertainty about the boolean value of p is given by a possibilistic evaluator for U, formally defined as:*

$$E_U : U^2 \rightarrow \tilde{\wp}(I) : (u, u') \mapsto E_U(u, u') = \{(T, \mu_{\tilde{p}}(T)), (F, \mu_{\tilde{p}}(F))\}$$

with $E_U(u, u') = E_U(u', u)$. Hereby, $\mu_{\tilde{p}}(T)$ represents the possibility that u and u' are co-referent and $\mu_{\tilde{p}}(F)$ represents the possibility that u and u' are not co-referent.

In Definition 1, symmetry is axiomatically required, because it is the only axiom of the equality relation in the entity world that can be translated to the evaluator. This is illustrated in Fig. 2, which shows the world of objects $(A, B, ...)$ describing entities $(X, Y, ...)$ in the real world.

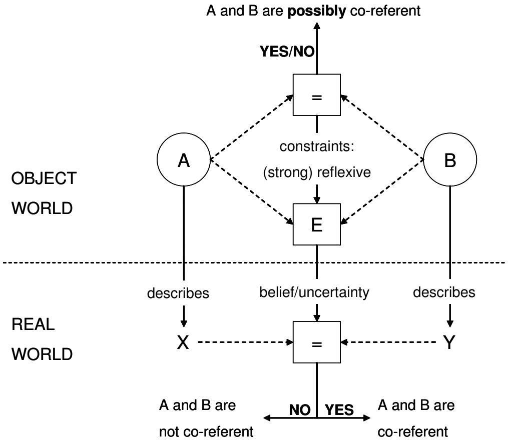

Figure 2. Difference between equality and co-reference

Note that in Definition 1, U represents the (sub)-object world. Co-reference is basically equality in the real world. However, (in)equality in the object world is merely a piece of evidence used by an evaluator E to describe the belief that the entities are equal. It follows that symmetry is the only axiom that any evaluator E should satisfy.

Object equality can be used by an evaluator E in two possible ways. If the constraint:

$$\forall (u, u') \in U^2 : u = u' \Rightarrow E_U(u, u') = \{(T, 1)\}$$

holds, E is a *reflexive* evaluator, stating that if two values from U are equal, they are certainly co-referent. A more strict constraint than reflexivity would be to assume that as soon as two values are not equal, it is to some extent possible that they are not co-referent. More specific, if the constraint:

$$\forall(u,u') \in U^2 : u = u' \Leftrightarrow E_U(u,u') = \{(T,1)\}$$

is satisfied, E is a *strong reflexive* evaluator.

A clear distinction is made between both constraints, which both have their use in applications. For example, let $\mathcal{C}$ denote the set of colors. If colors c_1 and c_2 are compared based on the human vision system, it follows that $E_{\mathcal{C}}(c_1,c_2) = \{(T,1)\}$ as soon as the human eye can not differentiate between the two colors. However, this does not ensure that both colors are equal. The actual problem is that the human vision system is not fully compatible with the equality relation in the universe $\mathcal{C}$, due to its limitations. Thus, in this case reflexivity is preferred over strong reflexivity. The same problem occurs when comparing real numbers that are stored on a computer and thus have a finite number of decimals.

In the scope of this paper, strong reflexivity is preferred because the equality relation is assumed to be known by the user (i.e. the user can decide upon equality). Therefor, non-equality always introduces some uncertainty. For example, assume two strings s=‘John Lennon’ and t=‘Jon Lennon’, it is not infeasible to state that $s \neq t$ implies that there is some (small) possibility that s and t are not co-referent. This example clarifies that, in many situations, two objects with small differences can be non co-referent, which is basically why co-reference is not compatible with similarity. Therefore, in what follows, evaluators are assumed to be strong reflexive, unless stated otherwise.

In some cases, it is possible to identify relations between propositions concerning co-reference. For instance, assume a proposition stating the co-reference of a and b, say $p_{a,b}$, and a proposition stating the same about b and c, say $p_{b,c}$. The uncertainty about the boolean truth values of these propositions is given by the PTVs $\tilde{p}_{a,b}$ and $\tilde{p}_{b,c}$. An interesting question is what can be derived about $p_{a,c}$, the proposition stating that a and c are co-referent. In the literature on similarity measures, such properties are often presented as transitivity, which is an implicit property of a similarity measure. In the possibilistic model for co-reference, such a direct transitivity is not present. Nevertheless, for a given evaluator, it might be possible and useful to derive a *conditional* possibility distribution over the domain $I = \{T, F\}$, say $\tilde{p}_{a,c}|\tilde{p}_{a,b}, \tilde{p}_{b,c}$ representing the uncertainty about the boolean value of $p_{a,c}$, given $\tilde{p}_{a,b}$ and $\tilde{p}_{b,c}$.

Let us start by describing the relations that we have. An indication that a and b are co-referent and an indication that b and c are co-referent, results in an indication for the co-reference of a and c. An indication that a and b (resp. b and c) are co-referent combined with an indication that b and c (resp. a and b) are *not* co-referent, yields an indication that a and c are not co-referent. Finally, an indication that a and b are not co-referent combined with an indication that b and c are not co-referent, tells us nothing about the co-reference of a and c. As an indicative measure for co-reference we consider necessity, which reflects *certainty* rather than possibility and is derived as follows:

$$\begin{aligned}&Nec(p = T) = 1 - Pos(p = F)\\&Nec(p = F) = 1 - Pos(p = T)\,.\end{aligned}$$

Based on these transformations and the following notations of conditional necessity:

$$\begin{aligned}&\mathcal{N}(p_{a,c} = T) = Nec(p_{a,c} = T|\tilde{p}_{a,b}, \tilde{p}_{b,c})\\&\mathcal{N}(p_{a,c} = F) = Nec(p_{a,c} = F|\tilde{p}_{a,b}, \tilde{p}_{b,c})\end{aligned}$$

the above descriptions of the (un)certainty relations between propositions are formalized as follows:

$$\begin{aligned}\mathcal{N}(p_{a,c} = T) &\geq Nec(p_{a,b} = T) \wedge Nec(p_{b,c} = T)\\\mathcal{N}(p_{a,c} = F) &\geq (Nec\,(p_{a,b} = T) \wedge Nec\,(p_{b,c} = F))\\&\quad \vee (Nec\,(p_{a,b} = F) \wedge Nec\,(p_{b,c} = T))\,.\end{aligned}$$

where the conjunction operator $\wedge$ is min and the disjunction operator $\vee$ is max. By adding the normalization condition of necessities $\min(\mathcal{N}(p_{a,c} = T), \mathcal{N}(p_{a,c} = F)) = 0$, the conditional necessities can be determined and hence the conditional possibilities by using the inverse transformations:

$$\begin{aligned}&\mu_{\tilde{p}_{a,c}|\tilde{p}_{a,b},\tilde{p}_{b,c}}(T) = Pos(p_{a,c} = T|\tilde{p}_{a,b}, \tilde{p}_{b,c}) = 1 - \mathcal{N}(p_{a,c} = F)\\&\mu_{\tilde{p}_{a,c}|\tilde{p}_{a,b},\tilde{p}_{b,c}}(F) = Pos(p_{a,c} = F|\tilde{p}_{a,b}, \tilde{p}_{b,c}) = 1 - \mathcal{N}(p_{a,c} = T)\end{aligned}$$

with $\tilde{p}_{a,c}|\tilde{p}_{a,b}, \tilde{p}_{b,c}$ the possibilistic truth value that represents the conditional uncertainty about the proposition $p_{a,c}$, given the uncertainty about $p_{a,b}$ and $p_{b,c}$. The conditional possibility distribution provides an upper bound for the uncertainty about proposition $p_{a,c}$, meaning that if additional information about this proposition becomes available, the resulting uncertainty must be smaller than or equal to the conditional uncertainty we had before the addition of information.

Table 1 contains some examples of derived conditional possibility distributions. The examples show how uncertainty about the basic propositions is contained in the conditional distribution. When there are indications that both basic properties are false, the conditional distribution will reflect complete uncertainty, just as is required.

The inference of a conditional possibility distribution can have important applications. Nevertheless, it is possible that $\tilde{p}_{a,c}|\tilde{p}_{a,b}, \tilde{p}_{b,c}$ is in *contradiction* with $E(a, c)$. Two PTVs, $\tilde{p}_1$ and $\tilde{p}_2$, concerning a proposition p are in contradiction if they indicate an opposite truth value as most possible:

$$(\mu_{\tilde{p}_1}(T) = 1 \wedge \mu_{\tilde{p}_2}(F) = 1) \vee (\mu_{\tilde{p}_1}(F) = 1 \wedge \mu_{\tilde{p}_2}(T) = 1)\ .$$

Table 1. Examples of conditional possibility distributions

$\tilde{p}_{a,b}$	$\tilde{p}_{b,c}$	$\tilde{p}_{a,c}\|\tilde{p}_{a,b},\tilde{p}_{b,c}$
(1,0)	(1,0)	(1,0)
(1,0)	(0,1)	(0,1)
(0,1)	(0,1)	(1,1)
(1,0)	(1,1)	(1,1)
(0,1)	(1,1)	(1,1)
(1,0.3)	(1,0.1)	(1,0.3)
(0.5,1)	(1,0.1)	(0.5,1)
(0.5,1)	(0.3,1)	(1,1)
(1,1)	(1,1)	(1,1)

A *consistent* evaluator is an evaluator for which $\tilde{p}_{a,c}|\tilde{p}_{a,b},\tilde{p}_{b,c}$ and $E(a,c)$ are never in contradiction. Typically, for any non-atomic data type such as sets and multisets, consistency between the conditional PTV and the actual evaluation is hard to guarantee. Given a universe U, construction of consistent evaluators is possible if U can be decomposed into *disjunct* subsets of U, say $U_1,...,U_m$, such that it is completely possible that two elements from the same U_i are co-referent and that it is completely possible that an element from U_i is not co-referent with any element from U_j, with $i \neq j$. This is formally stated by the following theorem:

Theorem 1 *Assume a universe U, E_U is a consistent evaluator if U can be decomposed in m sets $U_i \in \wp(U)$ such that:*

$$\forall i,j \in \{1,...,m\} : i \neq j \Rightarrow U_i \cap U_j = \emptyset$$
$$\forall i,j \in \{1,...,m\} : i \neq j \Rightarrow \forall (u,v) \in U_i \times U_j : \mu_{E_U(u,v)}(F) = 1$$
$$\forall i \in \{1,...,m\} : \forall (u,v) \in U_i^2 : \mu_{E_U(u,v)}(T) = 1\,.$$

This theorem is proved by a case study for three random values u, v and w and is therefore omitted here. Based on Theorem 1, some interesting consistent evaluators can be constructed. Assume a universe U and an equivalence relation R on U (reflexive, symmetric and transitive). Then E_U is consistent if it satisfies $\mu_{E_U(u,v)}(T) = 1 \Leftrightarrow (u\ R\ v)$ and $\mu_{E_U(u,v)}(F) = 1 \Leftrightarrow \neg(u\ R\ v)$. Practical examples of equivalence relations are equality, modulo and is-synonym-of. As a second example, assume U and a partial order relation P on U (reflexive, antisymmetric and transitive). Now E_U is consistent if it satisfies $\mu_{E_U(u,v)}(T) = 1 \Leftrightarrow ((u\ R\ v) \vee (v\ R\ u))$ and $\mu_{E_U(u,v)}(F) = 1 \Leftrightarrow \neg(u\ R\ v) \wedge \neg(v\ R\ u)$. Practical examples of partial order relations are subset for sets, substring and subsequence for strings and hierarchical relations in a geographic setting. As the remainder of the paper focuses on non-atomic data types, consistency is not assumed in what follows. In conclusion of this Section, a distinction is made between *semantical* evaluators and *syntactical* evaluators. On the one hand, semantical evaluators exploit semantical relations between values. For example, the strings "Manhattan" and "New York" are highly possibly co-referent, because both strings refer to a geographic region and one region is a part of the other. Hence, the semantics of both strings imply a high possibility of co-reference. Syntactical evaluators, on the other hand, are strictly based on a comparison of the syntactical structure of values. For example, the strings "John Lennon" and "Jon Lennon" are highly possibly co-referent due to a large syntactical proximity. In the remainder of the paper, the focus lies on syntactical evaluators.

5. Set evaluation

After introducing a very generic definition of co-reference evaluation, the first data type for which an evaluator is constructed, is the set-type. Sets are regularly used as data type to model unordered collections in which no duplicates are allowed. It is emphasized that a set (and consequently also a multiset in Section 6) is interpreted as a many valued attribute, rather than a complex object as is the case in Dubois and Prade (1982).

In the following, two approaches for set evaluation are provided. The first approach is an extension of regular set comparison techniques called *hard evaluation*, due to the use of element equality in calculations. The second approach, called *soft evaluation*, considers that elements themselves can be co-referent, while not equal. Uncertainty on element level is estimated using a low level evaluator. An injective element mapping between the two sets is constructed and the *unique* sequence of PTVs generated under this mapping is aggregated to a final result, representing the uncertainty about the co-reference of the sets.

5.1. Hard set evaluation

In a first approach some regular comparison techniques for sets are extended to $\tilde{\wp}(I)$. These comparison strategies calculate a result based on (well known) set functions (Cross and Sudkamp, 2002), such as union and intersection. After deriving sets by using such functions, an important step is the mapping of the derived sets to the unit interval. Dubois and Prade (1982) use *fuzzy measures* in this step of the comparison, which are defined as follows. Assume a universe U and two subsets of U, A and B. A *fuzzy measure* γ (Sugeno, 1974) is a mapping from $\wp(U)$ to $[0,1]$ satisfying $\gamma(\emptyset) = 0$, $\gamma(U) = 1$ and $A \subseteq B \Rightarrow \gamma(A) \leq \gamma(B)$. Our approach requires an estimation of the possibilities that two given sets are (not) co-referent. To obtain this, a couple of bipolar fuzzy measures is used to make a distinction between positive and negative information delivered by the results of set functions. In the context of sets, the positive information is contained in the elements shared by the sets and the negative information is contained in the elements that do not occur in the intersection of both sets. Hence, a formal way of defining a hard possibilistic set evaluator is:

Definition 2 (Hard Set Evaluator) *Assume a universe U. A hard set evaluator $E^h_{\wp(U)}$ is a strong reflexive evaluator, with:*

$$\mu_{E^h_{\wp(U)}(A,B)}(T) = s \cdot \frac{\gamma^T(A \cap B)}{\gamma^T(A \cup B)}$$
$$\mu_{E^h_{\wp(U)}(A,B)}(F) = s \cdot \frac{\gamma^F(A \Delta B)}{\gamma^F(A \cup B)}$$

where $A\Delta B = (A \cup B) \cap (\overline{A} \cup \overline{B})$ is the symmetrical difference of two sets and s is a scaling factor used to ensure normalization.

The fuzzy measures evaluate the relevance of the elements in a set. The possibility of co-reference is based on whether there is a significant difference in relevance between the elements in the intersection and the union. Similarly, the possibility of non co-reference is based on the difference in relevance between the symmetrical difference and the union. The relevance of an element being in the intersection might differ strongly from the relevance of that element being in the symmetric difference, which is why two functions γ^T and γ^F are used. More specifically, $x \in (A \cap B)$ might have a low relevance for the possibility that A and B are co-referent, while $x \in (A\Delta B)$ might be very relevant for the possibility that A and B are not co-referent. For both γ's, a simple example is $\gamma(A) = \frac{|A|}{|U|}$.

5.2. Soft set evaluation

5.2.1. Definition

Hard set evaluation is based on set functions that use the equality relation '=' on the universe of discourse. A second and more flexible approach generalizes the strict equality of elements to co-reference of elements. Thus, it is explicitly assumed that elements themselves can be co-referent without being equal. Measurement of the possibility of such lower level co-reference requires an evaluator E_U, which is assumed to be given. Obviously, E_U must satisfy Definition 1 and the assumption made in Section 4, i.e. strong reflexivity. Having two subsets of U, say A and B with $|A| \leq |B|$, the goal is to determine whether A and B are co-referent sets, by using E_U. The first step is creating an injection ι from A to B, thereby giving preference to couples of elements that are more possible to be co-referent, according to E_U. Next, ι implies a sequence of PTVs representing the knowledge of co-reference on element level. Finally, an aggregation operator for PTVs transforms the sequence to one PTV stating the possibility of co-reference of the sets. From these observations, a soft set evaluator is formally defined as:

Definition 3 (Soft Set Evaluator) *Assume a universe U and two subsets A and B with $|A| \leq |B|$. A soft set evaluator is a strong reflexive evaluator* *defined as:*

$$E^s_{\wp(U)}(A,B) = \phi(\kappa(A,B)) = \phi(\tilde{p}_1, ..., \tilde{p}_{|B|})$$

where $\kappa(A,B)$ *is a function that generates* $|B|$ *PTVs and* ϕ *is an aggregation function.*

In the following sections, both functions κ and ϕ are discussed.

5.2.2. Construction of ι and κ

The first important part in Definition 3 is the construction of a sequence of PTVs, implied by an injection between the elements of the sets. The construction of the injection can be split up into two basic steps.

In the first step, Algorithm 1 starts by mapping elements in $A \cap B$ to each other, which is justified by a corollary of strong reflexivity:

$$\forall (u,u') \in U^2 : u \neq u' \Rightarrow (\{(T,1)\} \,\tilde{=}\, E_U(u,u) \,\tilde{>}\, E_U(u,u'))$$

where $\tilde{=}$ and $\tilde{>}$ are generalized order relations as introduced in Section 2. Next, a matrix M of PTVs is created expressing uncertainty about the co-reference of elements from $A \backslash B$ and $B \backslash A$. Hereby, the functions $r(.)$ and $c(.)$ provide one-to-one mappings of elements from A and B to row and column indexes, which are natural numbers. Note that in Algorithm 1 the variables A and B are overwritten with their respective asymmetrical differences. Hence, in what follows it is assumed that, after execution of Algorithm 1, $A \cap B = \emptyset$, which simplifies our notations.

```
**Algorithm 1** Matrix generation
**Require:** $(A,B) \in \wp(U)^2 \wedge |A| \leq |B|$
**Ensure:** An $(|A\backslash B| \times |B\backslash A|)$-matrix $M$ of PTVs and partial $\iota$
  $C \leftarrow A \cap B$
  $\forall x \in C : \iota(x) = x$
  $A \leftarrow A \backslash C$
  $B \leftarrow B \backslash C$
  $\forall (a,b) \in A \times B : M[r(a), c(b)] = E_U(a,b)$
```

In the second step, we want to iteratively find the largest PTV $\tilde{p}$ in M, add the couple (x,y) to the mapping for which $E_U(x,y) = \tilde{p}$ and then remove the row and column corresponding to x and y.

This process is equivalent to Algorithm 2, which is explained in the following. For each row in M, the "largest" PTV is located with the understanding that comparison of PTVs is based on generalized order relations as explained in Section 2. This means that the largest PTV is the one with the lowest $\mu(F)$. If two rows, say r_1 and r_2, exist with the same location of the largest PTV,

Algorithm 2 Element mapping

```
Require: $(|A| \times |B|)$-matrix $M$ of PTVs
Ensure: Injective mapping $\iota$
  $\forall a \in A : m[r(a)] \leftarrow \arg\max_{b \in B} M[r(a), c(b)]$
  while $\exists (x, y) \in A^2 : x \neq y \wedge m[r(x)] = m[r(y)]$ do
    $\tilde{p}_1 \leftarrow M[r(x), c(m[r(x)])]$
    $\tilde{p}_2 \leftarrow M[r(y), c(m[r(y)])]$
    if $\tilde{p}_1 = \tilde{p}_2$ then
      $decision \leftarrow$ choose$(M[r(x)], M[r(y)])$
    end if
    if $\tilde{p}_1 \tilde{<} \tilde{p}_2 \vee decision = r(x)$ then
      $m[r(x)] \leftarrow$ search$(M[r(x)])$
    else
      $m[r(y)] \leftarrow$ search$(M[r(y)])$
    end if
  end while
  $\forall a \in A : \iota(a) = c^{-1}(m[r(a)])$
```

a *mapping conflict* is present. These conflicts are resolved one at a time as follows. If the conflicting PTVs are different, a sub procedure called **search** disables the position of the current maximum on the row with the smallest current maximum and searches for a new maximum for that row. In doing so, disabled positions are not taken into account. If the PTVs are equal on both rows, a sub procedure called **choose** will identify which row should be passed to procedure **search** for relocation of its maximum. The procedure **choose** selects the remaining PTVs from each row (i.e. enabled positions on the row) which results in two multisets of PTVs, say M_1 and M_2. Now, if $M_1 = M_2$, both rows contain the same PTVs on enabled positions. By convention, in this case we choose r_1. If $M_1 \subset M_2$ or $M_2 \subset M_1$, obviously the row corresponding to the largest multiset is chosen, because it contains all PTVs of the smaller multiset. If neither of these cases yield, we subtract $M_1 \cap M_2$ from both multisets and the multiset containing the largest PTV after subtraction is chosen. In this way the largest possible PTVs are left for future maximum relocation.

An example of the element mapping described by Algorithm 2 is shown in Fig. 3. For convenience, the PTVs are shown in couple notation. Further on, the current maxima are marked as the location where the PTV is underlined. When a position is disabled, the corresponding PTV is deleted. From step (a) to (b) the conflicts between the three rows are resolved. Because row 1 contains the largest PTV, the maxima on row 2 and 3 are relocated. From step (b) to (c) the conflict between row 2 and 3 is resolved on column 2. Both PTVs are equal, so **choose** will select row 2 for maximum relocation because $(1, 0.4) \tilde{>} (1, 0.8)$. In situation (c), no conflicts occur and the algorithm stops. It can be proved that Algorithm 2 always converges.

$$\text{(a)}\begin{bmatrix}\underline{(1,0.05)} & (1,0.8) & (0.3,1)\\ \underline{(1,0.2)} & (1,0.3) & (1,0.4)\\ \underline{(1,0.2)} & (1,0.3) & (1,0.8)\end{bmatrix}\quad\text{(b)}\begin{bmatrix}\underline{(1,0.05)} & (1,0.8) & (0.3,1)\\ & \underline{(1,0.3)} & (1,0.4)\\ & \underline{(1,0.3)} & (1,0.8)\end{bmatrix}\quad\text{(c)}\begin{bmatrix}\underline{(1,0.05)} & (1,0.8) & (0.3,1)\\ & & \underline{(1,0.4)}\\ & \underline{(1,0.3)} & (1,0.8)\end{bmatrix}$$

Figure 3. Example of element mapping

Theorem 2 *Given an $(r \times c)$-matrix M of possibilistic truth values with $r \leq c$, Algorithm 2 converges.*

Proof. Assume first $r = c$ and consider an $(r \times r)$-matrix M' with:

$$\forall (i,j) \in \{1,..,r\}^2 : M'[i,j] \in \{0,1\}$$

where a 1 in M' indicates the position of the current maximum on a row during the algorithm, which means that:

$$\forall i \in \{1,..,r\} : \sum_{j=1}^{r} M'[i,j] = 1$$

during the entire algorithm execution. After initial maximum selection, let n denote the number of columns for which $\sum_{i=1}^{r} M'[i,j] = 0$, with j being column index. By induction on n, the convergence can be proved. In the base case, $n = 1$, which means that there is one column, say k, for which $\sum_{i=1}^{r} M'[i,k] = 2$, and one column, say l, for which $\sum_{i=1}^{r} M'[i,l] = 0$. The algorithm will choose a row on which the current maximum is disabled and a new maximum is searched, which means that the elements of k sum up to 1 and the elements of a new column, say k', either sum up to 1 or 2. If $k' = l$, all columns sum up to 1 and the algorithm stops. If not, the previous situation remains with k' instead of k and one location disabled. Due to the finite number of locations in the matrix, the algorithm must eventually choose $k' = l$, which stops the algorithm. In the inductive case, the induction hypothesis states that for $n - 1 < r$ it is known that the algorithm will converge. If it can be proved for $n \leq r$, the inductive case is also proved. We have that there are n columns, for which elements sum up to 0 and $r - n$ columns, for which elements sum up to at least 1. The first set of columns is called A and the second set of columns is called B here. We have that $A \cap B = \emptyset$. Again, as the algorithm starts, it will change the position of 1 on some row, while disabling the previous position of 1 on this row. If the new column position of this 1 is on a column from A we can use the induction hypothesis to conclude the proof. If not, the new column position must be a column in B, which is the same situation as before the 1 has been replaced, with one position extra disabled. Due to the finite number of positions in the columns of B, it is certain that at a given time the algorithm must replace a 1 to a column from A, which, by use of the induction hypothesis, guarantees convergence. It can easily be seen that the foregoing proof also proves the case where $r < c$, because it simply increases the number of columns with elements that sum up to 0. ■

For each couple of elements from the created injection, E_U provides a PTV expressing uncertainty about the element co-reference. Note that $|B| - |A|$ elements from B have no image under the created injection, which implies that $|B| - |A|$ PTVs in the final generated sequence are $(0, 1)$. The generation of a sequence of PTVs is equivalent to Algorithm 3, which represents κ.

```
Algorithm 3 $\kappa$
Require: $A, B \subset U \wedge |A| \leq |B|$
Ensure: Sequence $seq$ of PTVs
  $\iota \leftarrow$Element_mapping(Matrix_generation(A,B))
  $\forall k \in \iota$ :add($seq, E_U(k)$)
  $\forall i \in \{1, .., |B| - |A|\}$ :add($seq, (0, 1)$)
```

Another important topic next to the convergence of Algorithm 2 is the uniqueness of image of κ.

Theorem 3 *Given an $(r \times c)$-matrix M of possibilistic truth values with $r \leq c$, the multiset of PTVs implied by the mapping provided by Algorithm 2, is unique, regardless of the order in which the mapping conflicts are resolved.*

Proof. The uniqueness of the resulting multiset of PTVs depends on two parts of the algorithm: **search** and **choose**. The first procedure locates the maximum PTV, based on a total order relation on PTVs. Hence, the choice of PTV is guaranteed to be consistent. The multiset of PTVs, in which the maximum is relocated, is determined by **choose**. In doing so, a partial order on multisets of PTVs is used. If two multisets are indistinguishable under this order (i.e. multisets are equal or one is a subset of the other), the choice does not affect the output of Algorithm 3. Consequently, when Algorithm 2 resolves conflicts between rows, the order in which the conflicts are resolved, does not affect the eventual sequence of PTVs. ■

After discussing the construction of an injective mapping between two sets, the complexity of the presented method is analyzed. Algorithm 1 constructs a matrix M, which has complexity $O(|A \backslash B| \cdot |B \backslash A| \cdot C(E_U))$ and is hence quadratic in terms of cardinality of the asymmetrical set differences. Next, Algorithm 2 constructs ι based on M. Assume an $(r \times c)$-matrix M. The initial search to locate maxima on each row is $O(r.c)$ on average, hence quadratic. Procedures **search** and **choose** are $O(c)$ on average and thus linear, which means the complexity of the iterations after initialization are linear. The number of iterations after initialization is $\frac{r.(r-1)}{4}$ on average. Algorithm 3 has linear complexity

Table 2. A mapping example

	c	d
a	(1,0.1)	(1,0.3)
b	(1,0.3)	(1,0.7)

$O(|B|)$. Hence, our method requires quadratic time to construct M, quadratic time to construct ι and linear time to construct the PTVs.

In conclusion, a brief comparison of the proposed algorithm with existing algorithms is given. As an alternative for the injective mapping, the *assignment algorithm* used to provide an optimal mapping in optimization problems could be considered. The reason why a new algorithm is introduced here is because the assignment algorithm (Silver, 1960) optimizes a global criterion, whereas here we are looking for optimization of a local criterion rather than a global one. By local criterion it is meant that the mapping must represent the couples of elements that are most possibly co-referent. For example, consider Table 2 which contains a matrix M with four PTVs (again shown in couple notation). Assume as global criterion for the selected set of PTVs S that $\min_S(\tilde{\bigwedge}_{s \in S} s)$. The global criterion would map b to c and a to d, whereas our algorithm maps a to c and b to d because it selects the couples that are most possible to be co-referent first. A consequence of this, which is also elaborated in the construction of the algorithm, is that equal elements across both sets should always be mapped onto each other. The reason why the proposed method is preferred is because the PTVs reflect (un)certainty and not some degree of preference. Indeed, when dealing with preferences, it makes sense to maximize the overall preference. However, the PTVs represent knowledge about reality, rather than preference, which leads to different semantics. In Table 2, with a and c very possibly co-referent, it would be semantically incorrect if these elements were not mapped to each other, just to maximize the possibility that the sets are co-referent. For this reason, global optimization in a setting of uncertainty or belief is not the correct solution. An alternative strategy for comparison of sets is given in Matthé et al. (2006). In contradiction to the algorithm for mapping proposed in Matthé et al. (2006), it is proved that the sequence of PTVs generated by the novel algorithm is always unique, which is considered a major benefit. In current literature, no algorithm that delivers a unique sequence of PTVs could be found.

5.2.3. Construction of ϕ

A second important part of Definition 3 is the aggregation function ϕ, which is an extension of a boolean connective for PTVs. The most obvious examples of such PTV-functions are (weighted) conjunctive and disjunctive operators (see De Cooman, 1995; De Tré and De Baets, 2003; Matthé, De Tré and Hallez, 2007). When using weighted operators it should be emphasized that the provided algorithm ensures only uniqueness of the image of κ, not the uniqueness of the injection ι. However, when using an ordered weighted aggregation function with a weight vector that is independent of the input values, only the generated PTVs determine output of ϕ. Therefore, the use of an ordered weighted extension of $\tilde{\wedge}$, defined in Section 2, is studied. The key benefit of such an ordered weighted conjunction, is the use of a quantifier. A regular conjunction of n propositions is true if *all* propositions are true, whereas ordered weighted conjunction of n propositions is true if L propositions are true. Here, L is a linguistic label like "most", "some","all",... Such a linguistic label is modeled (in a conjunctive setting) by a non-increasing quantifier function q_L, where for each $x \in dom(q)$, $1 - q_L(x)$ expresses the maximum compatibility of a fixed quantity $y < x$ with the linguistic quantity L. As an additional boundary condition, $\sup_{x \in dom(q_L)}(x)$ is fully compatible with L. Hence, the function q_L can provide n-dimensional weight vectors $\underline{w}$ by considering $n-1$ equal length intervals on $dom(q_L)$. Applied to the case of set comparison, $1 - w_i$ represents the necessity that the two sets are co-referent under the assumption of $i-1$ co-referent elements and $\max(|X|,|Y|) - i + 1$ non co-referent elements. In addition, if all elements are co-referent, so are the sets (due to reflexivity). Although a different approach is used, the idea of using fuzzy quantifiers in combination with PTVs, was first elaborated in Hallez et al. (2004). Functions to combine the necessities with PTVs are provided in De Tré and De Baets (2003), Matthé, De Tré and Hallez (2007), and Bronselaer and De Tré (2008). Ordered weighted conjunction can now be formally defined:

Definition 4 (Ordered weighted conjunction) *Let $\underline{\tilde{p}}$ denote a vector of PTVs and $\underline{w}$ a non-increasing vector of $[0,1]$-values weights, with $\max_i w_i = 1$. Ordered weighted conjunction is formally defined as:*

$$\tilde{\bigwedge}_{ow} : [0,1]^n \times \tilde{\wp}(I)^n \rightarrow \tilde{\wp}(I) : (\underline{w}, \underline{\tilde{p}}) \mapsto \tilde{\bigwedge}_{ow}(\underline{w}, \underline{\tilde{p}})$$

where

$$\tilde{\bigwedge}_{ow}(\underline{w}, \underline{\tilde{p}}) = \mathcal{T}_c^*(w_1, \tilde{q}_1)\tilde{\wedge}...\tilde{\wedge}\mathcal{T}_c^*(w_n, \tilde{q}_n)$$

and $\underline{\tilde{q}}$ is a vector containing all elements of $\underline{\tilde{p}}$ but with $\forall i, j \in \{1, ..., |\underline{\tilde{q}}|\} : i < j \Rightarrow \tilde{q}_i \tilde{\geq} \tilde{q}_j$.

Note that ordered weighted conjunction is a different operator than ordered weighted average (Yager, 1988) because the weights have different constraints and different semantics. When dealing with the comparison of sets, the fuzzy quantifier used can be parameterized by the cardinalities of both sets. Assume two sets X, Y with $|X| \leq |Y|$ and consider the following function:

$$q_{\alpha,\beta,\delta}(x) = \begin{cases} 1, & x < \alpha|X| \\ \delta, & x > |X| + \beta\left(|Y| - |X|\right) \\ 1 + \frac{(\delta-1)(x-\alpha|X|)}{(1-\alpha-\beta)|X|+\beta|Y|}, & else \end{cases}$$

with $(\alpha, \beta, \delta) \in [0,1]^3$. When comparing X and Y, the corresponding weight vector $\underline{w}$ can be calculated based on the parameterized quantifier as follows: $\forall i \in \{1, .., |Y|\} : w_i = q_{\alpha,\beta,\delta}(i)$. The function q has three parameters. The first two, α and β, determine the shape of the quantifier based on the set cardinalities. The third parameter δ determines the lower limit of the output of q and thus the upper limit of OWC in case the objects are non-equal. In most cases, δ is a number close to 0. If we consider a soft set evaluator that uses q to calculate $\underline{w}$ and $\delta = 0$, strong reflexivity of the evaluator is no longer satisfied.

A visual representation of the defined quantifier is shown in Fig. 4. The lower panel shows the special case where $\alpha = 1$ and $\beta = 0$. The parameters of q, i.e. $< \alpha, \beta, \delta >$ can be trained by applying a learning algorithm on a training set. In more advanced applications, the shape of the fuzzy quantifier can variate in function of the cardinality ratio $\frac{|X|}{|Y|} \in]0,1]$. If this ratio is close to 1, the relevance of β is less significant than when dealing with a small ratio. From this point of view, a more advanced learning algorithm learns a matrix P of parameters, where each row of P represents a parameter vector, indexed by an interval of cardinality ratios. In this way, different quantifiers are learned for different situations.

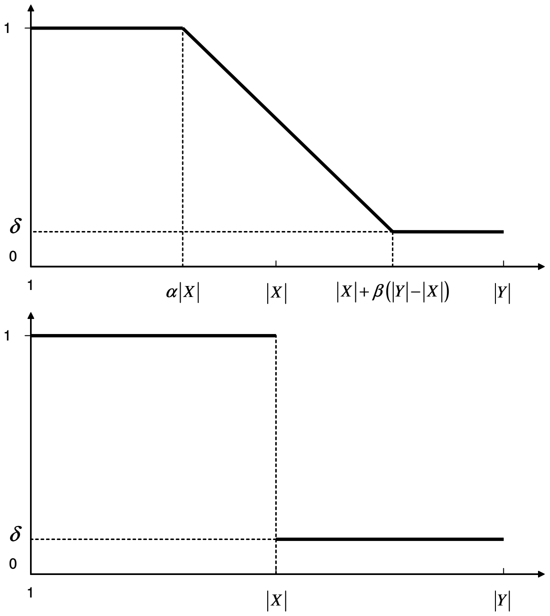

Figure 4. Parameterized fuzzy quantifiers for set comparison

An important question in such a strategy is the number of rows to consider in P. A first solution is to make a row for all the ratios encountered in the training set. However, for some ratios, the number of examples can be extremely low, so that insufficient samples are given to train the parameters. Further on, in cases where one quantifier is sufficient, the multiple quantifier approach might be a severe overkill. Finally, learning one quantifier for all different ratios, can cause overfitting of the model. Hence, a better approach would be to start by learning one quantifier. If the accuracy on the training set with one quantifier is insufficient, two quantifiers can be considered, where the first quantifier is trained for samples with a ratio in $[min, split]$ and the second quantifier is trained for samples with a ratio in $[split, 1]$. Hereby, *min* is the minimal cardinality ratio and *split* is chosen such that the number of samples available for each quantifier is approximately equal. This process continues until some stop criterion is met, for example if the accuracy does not increase significantly. When specifying the stop criterion it should be emphasized and taken into account that using too many quantifiers can result in overfitting.

Algorithm 4 Multiple quantifiers learner

```
Require: Training set $T$ with $< X, Y, coreferentFlag >$-samples and stop
  threshold $thr$
Ensure: $(r \times c)$-matrix containing $r$ parameter vectors
  $(i, \Delta) \leftarrow (1, 0)$
  repeat
    dim(P)=(i,c)
    $\underline{v} \leftarrow$ divideSamples$(T, i)$
    for $j = 1$ to $i$ do
      $P[i] \leftarrow$ optimalParameters$(\underline{v}_j)$
    end for
    $\Delta \leftarrow$ evaluate$(T, P) - \Delta$
    $i \leftarrow i + 1$
  until $\Delta < thr$
```

Algorithm 4 provides the pseudo code of the algorithm for learning multiple quantifiers q as defined before. Note that in this code the number of parameters for q is unspecified, which implies that any quantifier can be used, rather than just the one we introduced. Procedure **optimalParameters**$(\underline{v}_i)$ is an optimization strategy that finds optimal parameters for one quantifier q_i on a subset $\underline{v}_i$ of the original training set.

6. Multiset evaluation

After explaining some principles about set evaluation in the previous section, we will focus now on a well known extension of sets called *multisets* (Section 2). Yager pointed out that multisets can be useful in relational databases, more specifically - to extend some relational operators (Yager, 1986). In a more recent setting, the popular standard XML for the exchange of documents on the Internet, uses a tree structure in which the children of a node are a multiset of nodes, rather than a regular set of nodes. Further on, multisets can be used by a string evaluator if tokenization is used, resulting in a multiset of strings. In the light of these applications, it is useful to investigate the evaluation of multisets. As with sets in the previous Section, we will give two approaches for evaluation, quite similar to the case for regular sets. The first uses multiset functions and is called *hard multiset evaluation*, while the second one benefits from evaluation on element level and is called *soft multiset evaluation.*

6.1. Hard multiset evaluation

Elaborating the first approach similar to the case of regular sets requires the computation of union and intersection of multisets, which are introduced in Section 2, and the symmetrical difference $(A \ominus B) \cup (B \ominus A)$. An important problem to tackle is the use of fuzzy measures. Dubois and Prade (1982) suggested fuzzy set mappings onto the unit interval as an extension of regular fuzzy measures. However, given a universe U, the largest fuzzy set in terms of scalar cardinality that can be drawn from U, is U. Hence, formulating the condition $\gamma(U) = 1$, makes sense in the case of fuzzy sets, in contradiction to the case of multisets. It is easy to construct multisets that are larger than the original universe in terms of scalar cardinality of multisets. In fact, as the count function of multisets is mostly assumed to have an infinite image, it is impossible to construct the largest multiset. So, there is no use in defining direct extensions of fuzzy measures for multisets. It follows that *multiset measures* need to be formally defined to solve this issue. The need for such measures was pointed out by Rebai (1994) for the first time, who suggested the use of a reference multiset, which can be interpreted as a surrogate universe.

Definition 5 (Multiset Measure) *Assume a universe U and let $\mathcal{M}(U)$ be the set of all multisets drawn from U. For $A \in \mathcal{M}(U)$, a multiset measure based on A is a multiset function γ_A defined as:*

$$\gamma_A : \mathcal{A} \rightarrow [0,1] : B \mapsto \gamma_A(B)$$

with $\mathcal{A}$ being the set of all multisubsets of A and satisfying $\gamma_A(\emptyset) = 0$, $\gamma_A(A) = 1$ and $B \subset C \Rightarrow \gamma_A(B) \leq \gamma_A(C)$.

The key idea is that we will evaluate a multiset with respect to a supermultiset. Consider two multisets A and B for evaluation. As the relative universe can alter, it is sufficient to express the estimations relative to $A \cup B$, which serves as a limited universe that is *large enough.* Using such multiset measures, a hard possibilistic multiset evaluator is defined as follows.

Definition 6 (Hard multiset evaluator) *Assume the universe U. A hard multiset evaluator $E^h_{\mathcal{M}(U)}$ is a strong reflexive evaluator, with:*

$$\mu_{E^h_{\mathcal{M}(U)}(A,B)}(T) = s \cdot \gamma^T_{A \cup B}(A \cap B)$$
$$\mu_{E^h_{\mathcal{M}(U)}(A,B)}(F) = s \cdot \gamma^F_{A \cup B}((A \ominus B) \cup (B \ominus A))$$

where γ^T and γ^F are multiset measures and s is a scaling factor to ensure normalization.

Again, as with set evaluation, we allow different measures for estimation of possibility of T and F. Examples of multiset measures are:

$$\gamma_A(B) = \frac{\sum_{u \in U} C_B(u)}{\sum_{u \in U} C_A(u)} \text{ and } \gamma_A(B) = \frac{\sum_{u \in U} C_B(u) w_u}{\sum_{u \in U} C_A(u) w_u}.$$

The first measure simply compares the scalar cardinalities. The second uses a set of weights defined in the original universe U to assign a preference to elements. These weights must satisfy $\sum_{u \in U} w_u = 1$. It is also possible to extract a multiset measure from a regular fuzzy measure γ defined over the original universe U. There are several possible ways to define such extractions. For instance, consider the function:

$$g(A) = \max_{\bigoplus_i S_i = A} \left(\sum_i \gamma(S_i) \right)$$

with $A \in \mathcal{M}(U)$, $S_i \in \wp(U)$ and $\oplus$ the sum operator for multisets as defined in Section 2. This function g divides a multiset A into regular sets S_i that sum up to A and that result in a maximal evaluation sum. Another possibility is the function:

$$h(A) = \sum_{S \in D(A)} C_{D(A)}(S) \gamma(S)$$

with D being the decomposition function for multisets that transforms a multiset A into a minimal multiset of sets

$$D(A) = arg \min_{L = \{(i, S_i) | A = \bigoplus_i S_i\}} (|L|).$$

This later approach is more intuitive. Assume that the original universe of U is a multisubset of A, with $\min_{u \in U} C_A(u) = n$. This results in n "occurrences" of U in $D(A)$. As $\gamma(U) = 1$ by definition, the n occurrences of the universe are evaluated as $n\gamma(U) = n$. Using g or h we have:

$$\gamma_A(B) = \frac{g(B)}{g(A)} \text{ or } \gamma_A(B) = \frac{h(B)}{h(A)}$$

as multiset measures.

6.2. Soft multiset evaluation

As with sets, a second approach for multiset evaluation can be obtained by generalizing element equality to element co-reference. Most of the discussion of soft set evaluation can be translated directly to the case of multisets. The algorithms introduced in Section 5 do not require adaptation, but an important topic in the case of multisets is the low level evaluator used. A first option is to consider E_U, just as with the case of sets. It follows then, that each occurrence of an element is treated separately. A second approach is to consider $E_{U\times\mathbb{N}}$. In this case, groups of elements are considered as one entity and the possibility of co-reference of groups of elements is expressed. In this second approach, the number of occurrences of an element is considered an implicit property of the element, which is taken into account to determine the co-reference possibilities. The first approach is called *element based* evaluation, the second approach is called *support based* evaluation.

Definition 7 (Element based soft multiset evaluator)
Assume a universe U and two multisets A and B drawn from U, with $|A| \leq |B|$. An element based soft set evaluator is a strong reflexive evaluator defined as:

$$E^s_{\mathcal{M}(U)}(A,B) = \phi(\kappa(A,B)) = \phi(\tilde{p}_1, ..., \tilde{p}_{|B|})$$

with $\kappa(A,B) = \tilde{p}_1, ..., \tilde{p}_{|B|}$ being a function that generates $|B|$ PTVs and ϕ an aggregation function.

Here, construction of κ and ϕ is equivalent to the case of sets in Section 5.

Definition 8 (Support based soft multiset evaluator)
Assume a universe U and two multisets A and B drawn from U, with $|A| \leq |B|$. A support based soft set evaluator is a strong reflexive evaluator defined as:

$$E^s_{\mathcal{M}(U)}(A,B) = E^s_{\wp(U\times\mathbb{N})}(\widehat{A}, \widehat{B})$$

with

$$\forall X \in \mathcal{M}(U) : \widehat{X} = \{(u,n) | u \in U \wedge n = C_X(u)\}.$$

The element based evaluator is useful when dealing with situations where multiple occurrences of the same element can be assumed independent of each other. Such is the case when using a soft multiset evaluator in the context of estimating the co-reference of two strings that are tokenized into multisets of strings. The support based evaluator is useful when the multiple occurrences are related to each other. For example, when dealing with musical scores, each score contains lines for groups of musical instruments used in a particular song. In this case, it is more convenient to consider the groups of instruments as a whole.

7. Future work

The current paper discusses the concept of evaluators and defines such evaluators for sets and multisets. As mentioned before, finding co-referent multisets has a very interesting application in finding co-referent strings. Clearly, the string-datatype is used often in databases and OO-environments. As knowledge extraction from (textual) web sources into (semi)-structured data has become an important topic in current research, the importance of detecting co-referent strings has increased even more. Hence, applying the presented work in an application for detecting co-referent strings, is an urgent topic for future research. From this point of view, it can be beneficial to improve the presented work in some points. For example, it can be interesting to examine if the complexity of the mapping algorithm can be optimized if the low level evaluator E_U is consistent. Further on, having n possibilistic variables that are min-independent, ordered weighted conjunction has some interesting properties regarding calculation (Bronselaer and De Tré, 2008). Next to the complexity, the quantifier function could be studied more. The algorithm for dynamical adaptation of the used quantifier to a given situation is the first step in this direction. Finally, the study of evaluators should result in a study of the global possibilistic system for finding co-referent objects of arbitrary complexity.

8. Conclusion

In the presented research the object matching problem has been tackled from a possibilistic point of view, leading to a possibilistic solution for the co-reference problem. In order to further elaborate this model, a formal definition of evaluation operators is introduced in the domain of possibilistic truth values. These operators estimate possibilities concerning co-reference of (sub)-objects. As an application, evaluation operators for sets and multisets are presented, due to their practical applications in checking co-reference of many-valued attributes and strings. In both cases, two approaches are given. The first class of evaluators are an extension of existing work based on equality of elements and are called hard evaluators. The second and novel approach considers element co-reference, resulting in soft evaluators. An algorithm that creates an injection to be used by soft evaluators is given and the benefits of the algorithm are discussed. Next, it is shown how parameterized fuzzy quantifiers can be used in the aggregation step of the soft evaluators. A strategy to learn the optimal number of fuzzy quantifiers and their parameters is presented. In the case of multisets, two types of soft evaluators are distinguished: element-based, treating each element occurrence as a separate element and support based, treating the element count as a property of the element.

References

Bronselaer, A. and De Tré, G. (2008) Impact of [0,1]-valued weights and weighted aggregation operators for possibilistic truth values. In: *Proceedings of the NAFIPS 08 Congres.* IEEE, New York, 1–6.

Bronselaer, A., De Tré, G. and Hallez, A. (2007) Dynamic preference modeling in flexible object matching. In: *Proc. of the Eurofuse Workshop on New Trends in Preference Modeling.* University of Jaén, 191–195.

Cohen, W. (1998) Integration of heterogeneous databases without common domains using queries based on textual similarity. In: *Proc. of the ACM Sigmod Int. Conference of Management of Data.* ACM Press, 201–212.

Cross, V. and Sudkamp, T. (2002) *Similarity and Compatibility in Fuzzy Set Theory: Assessment and Applications.* Physica-Verlag.

De Cooman, G. (1995) Towards a possibilistic logic. In: Da Ruan, ed., *Fuzzy Set Theory and Advanced Mathematical Applications.* Kluwer Academic, Boston, 89–133.

Doan, A., Lu, Y., Lee, Y. and Han, J. (2003) Object matching for information integration: A profiler-based approach. In: *Proc. of the IJCAI-03 Workshop on Information Integration on the Web*, published online, 53–58.

Du Bois, N. (1969) A solution to the problem of linking multivariate documents. *American Statistical Association Journal* **64** (325), 163–174.

Du Bois, D. and Prade, H. (1982) *Recent Developments in Fuzzy Set and Possibility Theory.* Chapter: A unifying view of comparison indices in a fuzzy set-theoretic framework. Pergamon Press, 3–13.

Elmagarmid, A., Ipeirotis, P. and Verykios, V. (2007) Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering* **19** (1), 1–16.

Fellegi, I. and Sunter, A. (1969) A theory for record linkage. *American Statistical Association Journal* **64** (328), 1183–1210.

Hallez, A. and De Tré, G. (2007) A hierarchical approach to object comparison. In: *Proceedings of IFSA World Congres on Foundations of fuzzy logic and soft computing*, Cancun, Mexico. Springer, 191–198.

Hallez, A., De Tré, G., Verstraete, J. and Matthé, T. (2004) Application of fuzzy quantifiers on possibilistic truth values. In: *Proceedings of the Eurofuse Workshop on Data and Knowledge Engineering*, Warsaw, Poland. EXIT, Warsaw, 252–254.

Jaccard, P. (1908) Nouvelles recherches sur la distribution florale. *Bulletin de la Société de Vaud des Sciences Naturelles*, 44–223.

Jaro, M. (1989) Advances in record linking methodology as applied to the 1985 census of Tampa, Florida. *Journal of the American Statistical Society* **84** (406), 414–420.

Koyuncu, M. and Yazici, A. (2001) A fuzzy database and knowledge base environment for intelligent retrieval. In: *Proceedings of the IFSA/NAFIPS World Congress*, Vancouver, Canada. IEEE, 2311–2316.

Marín, N., Medina, J., Pons, O., Sanchez, D. and Villa, M. (2003) Complex object comparison in a fuzzy context. *Information and Software Technology* **45** 431–444.

Matthé, T., De Tré, G. and Hallez, A. (2007) Weighted conjunctive and disjunctive aggregation of possibilistic truth values. In: *Lecture Notes in Artificial Intelligence.* Springer, 171–180.

Matthé, T. et al. (2006) Similarity between multi-valued thesaurus attributes: Theory and application in multimedia systems. *Lecture Notes in Artificial Intelligence.* Springer, 331–342.

Prade, H. (1982) Possibility sets, fuzzy sets and their relation to lukasiewicz logic. In: *Proc. of the Int. Symposium on Multiple-Valued Logic.* Computer Society Press, 223–227.

Rebai, A. (1994) Canonical fuzzy bags and bag fuzzy measures as a basis for madm with mixed non cardinal data. *European Journal of Operational Research* **78**, 34–48.

Silver, R. (1960) An algorithm for the assignment problem. *Communications of the ACM* **3** (11), 605–606.

Sugeno, M. (1974) *Theory of fuzzy integrals and its applications.* PhD thesis, Tokyo Institute of Technology.

Tejada, S., Knoblock, C. and Minton, S. (2001) Learning object identification rules for information integration. *Inf. Systems* **26** (8), 607–633.

De Tré, G. and De Baets, B. (2003) Aggregating constraint satisfaction degrees expressed by possibilistic truth values. *IEEE Transactions of Fuzzy Systems* **11** (3), 361–368.

Tversky, A. (1977) Features of similarities. *Psychological Rev.* **84**, 327–352.

Wang, Y. and Madnick, S. (1989) The inter-database instance identification problem in integrating autonomous systems. In: *Proc. of the Fifth IEEE International Conference on Data Engineering.* IEEE Computer Society, 46–55.

Weis, M. and Naumann, F. (2004) Detecting duplicate objects in xml documents. In: *Proceedings of the 2004 International Workshop on Information Quality in Information Systems*, Paris, France. ACM Press, 10–19.

Winkler, W. (1993) Improved decision rules in the Fellegi-Sunter model of record linkage. *Techn. Report RR93/12, Statistical Research Report Series.*

Winkler, W. (2006) Overview of record linkage and current research directions. *Technical Report RRS2006/02, Statistical Research Report Series.*

Yager, R. (1986) On the theory of bags. *International Journal of General Systems* **13** (1), 23–27.

Yager, R. (1988) On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Transactions on Systems, Man and Cybernetics* **18** (1), 183–190.

Zadeh, L. (1978) Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems* (1), 3–28.