Measuring the quality of uncertain information using possibilistic logic

Anthony Hunter[1] and Weiru Liu[2]

[1]Department of Computer Science
University College London
Gower Street, London WC1E 6BT, UK
[2]School of Computer Science
Queen's University Belfast
Belfast, Co Antrim BT7 1NN, UK

Abstract. In previous papers, we have presented a framework for merging structured information in XML involving uncertainty in the form of probabilities, degrees of beliefs and necessity measures [HL04,HL05a,HL05b]. In this paper, we focus on the quality of uncertain information before merging. We first provide two definitions for measuring information quality of individually inconsistent possibilistic XML documents, and they complement the commonly used concept of inconsistency degree. These definitions enable us to identify if an XML document is of *good* or *lower* quality when it is inconsistent, as well as enable us to differentiate between documents that have the same degree of inconsistency. We then propose a more general method to measure the quality of an inconsistent possibilistic XML document in terms of a pair of coherence measures.

1 Introduction

With the increasing use of XML for representing information on the Web, the need for modelling uncertainty in the information has emerged. A probabilistic approach is taken in [NJ02] which provides an XML structure to model and reason with probabilistic values attached to different levels of tags in a single XML document. The final probability of the value of a specific tag is calculated as multiple conditional probabilities on its ancesters' tags. In another approach, [KKA05] probability values are also attached to tags, but require that the probabilities of a set of values associated with a single tag must sum to 1.0, a condition that was not required in [NJ02]. A simple merging method is provided to integrate two probabilitsic XML trees in [KKA05], whilst [NJ02] did not consider multiple XML documents. Both approaches are strongly rooted in relational databases and many operators, including queries are extensions of operations for probabilistic relational databases.

In contrast, method of modelling, reasoning, and merging XML documents with uncertain information in our research ([HL04,HL05a,HL05b]) concerns information within the logical fusion framework [HS04]. We use probability theory Dempster-Shafer theory, and possibility theory to model different types of uncertainty, as well as provide integration and aggregation mechanisms to merge multiple XML documents.

However, none of the research above has considered assessing the quality of uncertain information modelled in an XML document. In this paper, we focus on XML documents where uncertainties are modelled by necessity measures and attempt to assess the quality of uncertain information when inconsistency occurs. We will proceed as follows: (Sec.2) we present formal definitions for possibilistic information in structured reports (a form of XML document). (Sec.3) we propose two definitions to identify a *good* quality structured report from a *lower* quality structured report when they both have the same degree of inconsistency. We also discuss how coherence measures can be used to measure the quality of an inconsistent structured report when it does not fall into either *good* or *lower* quality categoties. Sect. 4 concludes the paper.

2 Preliminaries

We now provide basic definitions for structured reports, for possibility theory, and for representing uncertain information in terms of necessity measures in structured reports.

2.1 Structured reports

We use XML to represent structured reports. So each structured report is an XML document, but not vice versa. If φ is a tagname (i.e an element name), and ϕ is a textentry, then $\langle\varphi\rangle\phi\langle/\varphi\rangle$ is a structured report. If φ is a tagname (i.e an element name), ϕ is a textentry, θ is an attribute name, and κ is an attribute value, then $\langle\varphi\ \theta = \kappa\rangle\phi\langle/\varphi\rangle$ is a structured report. If φ is an tagname and $\sigma_1, ..., \sigma_n$ are structured report, then $\langle\varphi\rangle\sigma_1 ... \sigma_n\langle/\varphi\rangle$ is a structured report.

Each structured report is isomorphic with a ground term of classical logic. This isomorphism is defined inductively as follows: (1) If $\langle\varphi\rangle\phi\langle/\varphi\rangle$ is a structured report, where ϕ is a textentry, then $\varphi(\phi)$ is a term that is isomorphic; (2) If $\langle\varphi\ \theta = \kappa\rangle\phi\langle/\varphi\rangle$ is a structured report, where ϕ is a textentry, then $\varphi(\phi, \kappa)$ is a term that is isomorphic; and (3) If $\langle\varphi\rangle\phi_1..\phi_n\langle/\varphi\rangle$ is a structured report, and ϕ'_1 is a term that is isomorphic with ϕ_1, ..., and ϕ'_n is a term that is isomorphic with ϕ_n, then $\varphi(\phi'_1, .., \phi'_n)$ is a term that is isomorphic.

2.2 Possibility theory

Let Ω be a frame of discernment containing all the distinctive and exhaustive solutions to a question. A possibility measure and a necessity measure in possibility theory [DP88,SDK95,BDP97], denoted Π and N respectively, are functions from $\wp(\Omega)$ to $[0, 1]$ such that $\Pi(\wp(\Omega)) = 1$, $\Pi(\emptyset) = 0$, and $N(A) = 1 - \Pi(\bar{A})$.

$\Pi(A)$, the degree of possibility assigned to A, estimates to what extent the true event is possibly in A, and $N(A)$, the degree of necessity assigned to A, evaluates to what extent the true event is believed to be in A.

Both possibility measure and necessity measure can be derived from a more elementary assignment, $\pi : \Omega \rightarrow [0, 1]$, which is referred to as a **possibility distribution**. The relationship between Π and π is

$$\Pi(A) = \mathsf{max}(\{\pi(\phi) | \phi \in A\})$$

⟨possibility⟩
⟨ness value = "0.5"⟩
⟨nessitem⟩8°C⟨/nessitem⟩
⟨nessitem⟩10°C⟨/nessitem⟩
⟨/ness⟩
⟨ness value = "0.8"⟩
⟨nessitem⟩12°C⟨/nessitem⟩
⟨/ness⟩
⟨/possibility⟩

Fig. 1. A possibility-valid component (a PVC).

which satisfies $\Pi(A \cup B) = \max(\Pi(A), \Pi(B))$. The usual condition associated with π is there exists ϕ_0 such that $\pi(\phi_0) = 1$, and in which case π is said to be *normal*.

2.3 Representing uncertain information in strucured reports

We extend the definitions for structured reports to represent uncertainty.

Definition 1. *The structured report* $\langle\texttt{possibility}\rangle\sigma_1, .., \sigma_n\langle/\texttt{possibility}\rangle$ *is called a* **possibility-valid component** *(a PVC) iff for each* $\sigma_i \in \{\sigma_1, .., \sigma_n\}$, σ_i *is of the form*

$$\langle\texttt{ness value} = \kappa\rangle\sigma_1^i, ..., \sigma_m^i\langle/\texttt{ness}\rangle$$

and for each $\sigma_j^i \in \{\sigma_1^i, .., \sigma_n^i\}$, σ_j^i *is of the form* $\langle\texttt{nessitem}\rangle\phi\langle/\texttt{nessitem}\rangle$ *and* $\kappa \in [0, 1]$, *and* ϕ *is a textentry.*

In possibility theory, both a degree of possibility (from Π) and a degree of necessity (from N) can be assigned to subsets of a set of possible values. In possibilistic logic, a weighted formula (p, a) implies that the weight a attached to formula p is interpreted as a lower bound on the degree of necessity $N(p)$ (with $N(p)$ being seen as a degree of belief on p) [BDP97,BDKP00]. In the context of this paper, a weight κ_i attached to a subset $\{\phi_l^1, ..., \phi_l^r\}$ is equally interpreted as a lower bound on the degree of necessity of $\{\phi_l^1, ..., \phi_l^r\}$. This also explains why we use tagname "ness" instead of "poss".

The textentries in a PVC are elements of a pre-defined set containing mutually exclusive and exhaustive values for the related tagname. A structured report involving uncertain information with necessity measures should satisfy the following constraints.

Definition 2. *Let* $\langle\texttt{possibility}\rangle\sigma_1, .., \sigma_n\langle/\texttt{possibility}\rangle$ *be a PVC, and let* $\sigma_i \in \{\sigma_1, .., \sigma_n\}$ *be of the form* $\langle\texttt{ness value} = \kappa_i\rangle\sigma_i^1, .., \sigma_i^p\langle/\texttt{ness}\rangle$, *and let* σ_i^k *be of the form* $\langle\texttt{nessitem}\rangle\phi_i^k\langle/\texttt{nessitem}\rangle$ *for* $1 \leq k \leq p$. *This component adheres to the* **necessity measure constraint** *in possibility theory iff the following conditions hold: (1)* $\kappa_i \in [0, 1]$ *(2) for all* i,j, *if* $1 \leq i \leq n$ *and* $1 \leq j \leq n$ *and* $i \neq j$, *then* $\{\phi_i^1, ..., \phi_i^p\} \neq \{\phi_j^1, ..., \phi_j^q\}$.

In contrast to situations in possibilistic logic where a possibilistic knowledge base can have both (p, a_1) and (p, a_2) where $a_1 \neq a_2$ are two degrees of necessity (each of which can be seen as a degree of belief) on the same logical sentence. In this case, (p, a_1) subsumes (p, a_2) when $a_1 > a_2$. Definition 2 restricts XML representation to the case where for each subset, there is only one degree of necessity associated with it in structured reports. This will reduce unnecessary XML segments in structured reports.

2.4 From necessity measures to possibility distributions

A PVC usually specifies a partial necessity measure. Here we recover the possibility distribution associated with this necessity measure using the minimum specificity principle. Let a PVC be $\langle \texttt{possibility} \rangle \sigma_1, \ldots, \sigma_p \langle / \texttt{possibility} \rangle$ where $\sigma_i \in \{\sigma_1, \ldots, \sigma_p\}$ is of the form $\langle \texttt{ness value} = \kappa_i \rangle \psi_i \langle / \texttt{ness} \rangle$ and ψ_i is of the form

$$\langle \texttt{nessitem} \rangle \phi_{i_1} \langle / \texttt{nessitem} \rangle \cdots \langle \texttt{nessitem} \rangle \phi_{i_x} \langle / \texttt{nessitem} \rangle$$

We denote the frame associated with a PVC as $\Omega = \{\phi_1, \ldots, \phi_n\}$, and also let $\psi_i = \{\phi_{i_1}, \ldots, \phi_{i_x}\}$ in order to make the subsequent description simpler. In this way, a PVC can be viewed as consisting of a finite set of weighted subsets of Ω, $\{(\psi_i, \kappa_i), i = 1, \ldots, p\}$, where κ_i is interpreted as a lower bound on the degree of necessity $N(\psi_i)$. This representation is consistent with notations in [DP87a] and analogous with notations in possibilistic knowledge bases using possibilistic logic, where uncertain knowledge is represented as a set of weighted formulae, $\{(p_i, a_i), i = 1, \ldots, n\}$. A subset ψ_i and formula p_i are thought to be equivalent if p_i is defined as $p_i = \vee q_j$, where q_j stands for "$\phi_j \in \psi_i$ is true". Therefore, when one of the elements in ψ_i is definitely true, formula p_i is definitely true as well.

Given a PVC, there is normally a family of possibility distributions associated with it and each of the distributions satisfying the condition $1 - \max\{\pi(\phi) | \phi \in \bar{\psi}_i\} \geq \kappa_i$. A common method to select one of the compatible possibility distributions is to use the **minimum specificity principle** [DP87a]. The minimum specificity principle allocates the greatest possibility degrees in agreement with the constraints $N(\psi_i) \geq \kappa_i$. This possibility distribution always exists [DP87a,BDP97] and is characterized as

$$\forall \phi \in \Omega, \pi(\phi) = \begin{cases} \min\{1 - \kappa_i | \phi \notin \psi_i\} & \text{when } \exists\ \psi_i \text{ such that } \phi \notin \psi_i \\ \quad = 1 - \max\{\kappa_i | \phi \notin \psi_i\} & \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

Definition 3. *Let a PVC be* $\langle \texttt{possibility} \rangle \sigma_1, \ldots, \sigma_p \langle / \texttt{possibility} \rangle$ *where (1)* $\sigma_i \in \{\sigma_1, \ldots, \sigma_p\}$ *is in the form* $\langle \texttt{ness value} = \kappa_i \rangle \psi_i \langle / \texttt{ness} \rangle$*; and (2)* ψ_i *is of the form* $\langle \texttt{nessitem} \rangle \phi_{i_1} \langle / \texttt{nessitem} \rangle \cdots \langle \texttt{nessitem} \rangle \phi_{i_x} \langle / \texttt{nessitem} \rangle$ *and the set of weighted subsets is* $\{(\psi_i, \kappa_i), i = 1, \ldots, p\}$*. Let* **the possibility distribution obtained using the minimum specificity principle** *be* $\pi : \Omega \to [0, 1]$*, where for each* $\phi \in \Omega$*,* $\pi(\phi) = 1 - \nu$ *and*

$$\nu = \begin{cases} \max\{\kappa_1, \kappa_2, \ldots, \kappa_t\}\ \phi \notin \psi_j, j = 1, 2, \ldots, t\ \ (\textit{ where } p \geq t > 0) \\ 0 \qquad\qquad\qquad\qquad\qquad \textit{otherwise} \end{cases}$$

⟨possibility⟩
  ⟨ness value = "0.2"⟩
    ⟨nessitem⟩ϕ_1⟨/nessitem⟩
    ⟨nessitem⟩ϕ_2⟨/nessitem⟩
  ⟨/ness⟩
  ⟨ness value = "0.3"⟩
    ⟨nessitem⟩ϕ_3⟨/nessitem⟩
  ⟨/ness⟩
⟨/possibility⟩

⟨possibility⟩
  ⟨ness value = "0.2"⟩
    ⟨nessitem⟩ϕ_1⟨/nessitem⟩
    ⟨nessitem⟩ϕ_2⟨/nessitem⟩
  ⟨/ness⟩
  ⟨ness value = "0.3"⟩
    ⟨nessitem⟩ϕ_2⟨/nessitem⟩
    ⟨nessitem⟩ϕ_3⟨/nessitem⟩
  ⟨/ness⟩
⟨/possibility⟩

Fig. 2. Possibility-valid components (PVCs) ($\Omega = \{\phi_1, \phi_2, \phi_3, \phi_4\}$).

Example 1. The possibility distribitions π_1 and π_2 below are obtained from the left and right PVCs in Figure 2 respectively using Eq (1).

$$\pi_1(\phi_1) = 0.7, \pi_1(\phi_2) = 0.7, \pi_1(\phi_3) = 0.8, \pi_1(\phi_4) = 0.7$$
$$\pi_2(\phi_1) = 0.7, \pi_2(\phi_2) = 1, \pi_2(\phi_3) = 0.8, \pi_2(\phi_4) = 0.7$$

3 Quality of uncertain information with inconsistency

3.1 Inconsistency degree

A possibility distribution is not normal if $\forall\phi, \pi(\phi) < 1$. The value $1 - \max_{\phi\in\Omega}\pi(\phi)$ is called **the degree of inconsistency** of the original PVC and is denoted as $\mathsf{Inc}(K)$ where K is the knowledge associated with the possibility distribution of the PVC. For instance, in Example 1, the PVC on the left is inconsistent since $\forall\phi, \pi(\phi) < 1$, whilst the right one is consistent, because $1 - \max_{\phi\in\Omega}(\pi_2(\phi)) = 0$.

Proposition 1. *Let $\{(\psi_i, a_i),\ i = 1, \ldots, p\}$ be weighted subsets of Ω and specified in a PVC with respect to frame of discernment Ω. This PVC is* **consistent** *iff $\cap_i\psi_i \neq \emptyset$, otherwise the PVC is* **inconsistent**.

Example 2. Consider the two PVCs in Figure 3. The possibility distributions from them using Equation (1) are

$$\pi_1(\phi_1) = 0.7, \pi_1(\phi_2) = 0.7, \pi_1(\phi_3) = 0.7, \pi_1(\phi_4) = 0.7, \pi_1(\phi_5) = 0.7, \pi_1(\phi_6) = 0.7$$
$$\pi_2(\phi_1) = 0.7, \pi_2(\phi_2) = 0.7, \pi_2(\phi_3) = 0.7, \pi_2(\phi_4) = 0.7, \pi_2(\phi_5) = 0.7, \pi_2(\phi_6) = 0.7$$

The degrees of inconsistencies of the two PVCs are the same, $1-\max_{\phi\in\Omega}(\pi_1(\phi)) = 0.3$ and $1 - \max_{\phi\in\Omega}(\pi_2(\phi)) = 0.3$. However, if we examine the structure of the weighted subsets ψ_i^1 and ψ_j^2 in detail, we will find that the right-hand side PVC is more coherent than the left one, since there is a significant overlap among the subsets ψ_j^2 in this PVC. While any two subsets in the first PVC have no common elements. This observation leads to the definitions below that further differentiates between *good* and *lower* qualities of an inconsistent PVC.

⟨possibility⟩
⟨ness value = “0.2”⟩
⟨nessitem⟩ϕ_1⟨/nessitem⟩
⟨nessitem⟩ϕ_2⟨/nessitem⟩
⟨/ness⟩
⟨ness value = “0.3”⟩
⟨nessitem⟩ϕ_3⟨/nessitem⟩
⟨nessitem⟩ϕ_4⟨/nessitem⟩
⟨/ness⟩
⟨ness value = “0.2”⟩
⟨nessitem⟩ϕ_5⟨/nessitem⟩
⟨/ness⟩
⟨ness value = “0.3”⟩
⟨nessitem⟩ϕ_6⟨/nessitem⟩
⟨/ness⟩
⟨/possibility⟩

⟨possibility⟩
⟨ness value = “0.2”⟩
⟨nessitem⟩ϕ_1⟨/nessitem⟩
⟨nessitem⟩ϕ_2⟨/nessitem⟩
⟨/ness⟩
⟨ness value = “0.3”⟩
⟨nessitem⟩ϕ_2⟨/nessitem⟩
⟨nessitem⟩ϕ_3⟨/nessitem⟩
⟨/ness⟩
⟨ness value = “0.3”⟩
⟨nessitem⟩ϕ_2⟨/nessitem⟩
⟨nessitem⟩ϕ_4⟨/nessitem⟩
⟨/ness⟩
⟨ness value = “0.3”⟩
⟨nessitem⟩ϕ_4⟨/nessitem⟩
⟨nessitem⟩ϕ_5⟨/nessitem⟩
⟨/ness⟩
⟨/possibility⟩

Fig. 3. PVCs ($\Omega = \{\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6\}$).

Definition 4. *Let* ⟨possibility⟩$\sigma_1, \ldots, \sigma_p$⟨/possibility⟩ *be PVC where (1)* $\sigma_i \in \{\sigma_1, \ldots, \sigma_p\}$ *is in the form* ⟨ness value $= \kappa_i$⟩ψ_i⟨/ness⟩*; (2)* ψ_i *is of the form* ⟨nessitem⟩ϕ_{i_1}⟨/nessitem⟩ $\cdots$ ⟨nessitem⟩ϕ_{i_x}⟨/nessitem⟩ *and the corresponding set of weighted subsets be* $\{(\psi_i, \kappa_i), i = 1, \ldots, p\}$*. This PVC is said to be* **inconsistent with good quality**, *if there exists a* ψ_j*, called a* **separable element**, *such that*

$$(\bigcap_{i=1, i\neq j}^{p} \psi_i) \neq \emptyset \text{ and } \bigcap_{i=1}^{p} \psi_i = \emptyset \tag{2}$$

Given a PVC, there can be several separable elements ψ_j satisfying this definition. This definition identifies those PVCs each of which would have a normal possibility distribution recovered from it when the identified subset ψ_j is deleted from the PVC. As a consequence, we provide an addition normalization rule that is best suited for this type of PVCs. We assign the maximum degree of possibility to the elements that have appeared in all but one subset in a PVC which also have the highest possibility value prior to normalization.

$$\pi_{n_4}(\phi) = \begin{cases} 1 & \phi \in (\bigcap_{i=1}^{p} \psi_i), \psi_i \neq \psi_j, \psi_j \text{is a separable element in Def. 4} \\ & \text{s.t. if} \exists \phi_l \in (\bigcap_{i=1}^{p} \psi_i), \psi_i \neq \psi_l, \text{is a separable element} \\ & \text{in Def. 4 then } \pi(\phi) > \pi(\phi_l) \\ \pi(\phi) & \texttt{otherwise} \end{cases} \tag{3}$$

When there are several elements $\phi_i, ..., \phi_j$ satisfying Eq (3) and they all have the same degree of possibility distribution, e.g., $\pi(\phi_i) = \pi(\phi_j)$, then we arbitrarily choose one of them to normalize.

This rule harnesses the 2nd of the three commonly used normalization rule as reviewed in [BDP97]:

$$\pi_{n_1}(\phi) = \frac{\pi(\phi)}{\mathsf{max}\{\pi(\phi_i)\}} \tag{4}$$

$$\pi_{n_2}(\phi) = \begin{cases} 1 & \texttt{if}\pi(\phi) = \mathsf{max}\{\pi(\phi_i)\} \\ \pi(\phi) & \texttt{otherwise} \end{cases} \tag{5}$$

$$\pi_{n_3}(\phi) = \pi(\phi) + (1 - \mathsf{max}\{\pi(\phi_i)\}) \tag{6}$$

As we can see, no matter which rule among this three we choose to apply, the normalized possibility distributions for the two PVCs in Fig. 3 are both reduced to a uniform distribution, e.g., for every $\phi \in \Omega$, $\pi(\phi) = 1$. However, using the new normalization rule, the right-hand side PVC in Fig. 3 has a normalized possibility distribution $\pi'_2(\phi_1) = 0.7, \pi'_2(\phi_2) = 1.0, \pi'_2(\phi_3) = 0.7, \pi'_2(\phi_4) = 0.7, \pi'_2(\phi_5) = 0.7, \pi'_2(\phi_6) = 0.7.$ which assigns 1 to element ϕ_2 only. This rule produces a better normalized possibility distribution than all the other three rules.

A separable element ψ_j can be disjoint with the rest of the weighted subsets completely or it can share common elements with some weighted subsets. This leads to the following definition.

Definition 5. *Let K be a PVC with a set of weighted subsets $S = \{(\psi_i, \kappa_i), i = 1, \ldots, p\}$. ψ is called an* **isolated separable element** *if the following condition holds*

$$\forall(\psi_i, \kappa_i) \in S, \psi_i \cap \psi = \emptyset \ \textit{ when } \ \psi_i \neq \psi.$$

Lemma 1. *Let K be a PVC which is inconsistent with good quality, if K has an isolated separable element ψ, then ψ is the only separable element.*

Proposition 2. *Let K be a PVC which is inconsistent with good quality and it has an isolated separable element ψ_t where $\kappa_t \geq \kappa_i$ for all other weighted subsets (ψ_i, κ_i) for $i = 1, ..., p, i \neq t$, then*

$$\mathsf{Inc}(K) = \mathsf{max}(\kappa_i | i \neq t)$$

Definition 6. *Let $\langle\texttt{possibility}\rangle\sigma_1, \ldots, \sigma_p\langle/\texttt{possibility}\rangle$ be a PVC where (1) $\sigma_i \in \{\sigma_1, \ldots, \sigma_p\}$ is in the form $\langle\texttt{ness value} = \kappa_i\rangle\psi_i\langle/\texttt{ness}\rangle$; and (2) ψ_i is of the form $\langle\texttt{nessitem}\rangle\phi_{i_1}\langle/\texttt{nessitem}\rangle \cdots \langle\texttt{nessitem}\rangle\phi_{i_x}\langle/\texttt{nessitem}\rangle$ and the corresponding set of weighted subsets be $\{(\psi_i, \kappa_i), i = 1, \ldots, p\}$. This PVC is said to be* **inconsistent with lower quality**, *if for every pair (ψ_i, ψ_j), $\psi_i \cap \psi_j = \emptyset$, when $\psi_i \neq \psi_j$.*

It is easy to see that every weighted subset in such a PVC is an isolated separatable element.

Proposition 3. *Let K be a PVC which is inconsistent with lower quality. Then the degree of inconsistency of this PVC is as follows where $\mathsf{max}^{\mathsf{2nd}}$ is a function that selects the 2nd largest value in a set of values $(\kappa_1, ..., \kappa_p)$.*

$$\mathsf{Inc}(K) = \mathsf{max}^{\mathsf{2nd}}\{\kappa_i | (\psi_i, \kappa_i)\}$$

However, these two definitions only describe the two extreme situations where in one case, all but one subset share some common elements, whlist in the other, all the subsets are separated from each other. In reality, many PVCs do not fall into these categories. We address this next.

3.2 Coherence measures

Since an inconsistency degree alone is not sufficient to reflect the quality of an inconsistent PVC in terms of the coherence of its weighted subsets, we propose a method to further assess the quality of such a PVC.

In [DKP03], a coherence function which extends the coherence measure in [Hun02] was proposed to measure the quality of a possibilistic knowledge base when inconsistency exists. We adapt this function here in terms of weighted subsets and provide our coherence measures of an inconsistent PVC.

Definition 7. *Let K be a PVC.*

$$\mathsf{OpinionBase}(K) = \{(\psi_i, \kappa_i) | \textit{ such that } (\psi_i, \kappa_i) \textit{ is a weighted subset of } K\}$$

$$\mathsf{ConflictBase}(K) = \{(\psi_i, \kappa_i) \in \mathsf{OpinionBase}(K) | \exists (\psi_{i_t}, \kappa_{i_t}) \in \mathsf{OpinionBase}(K), \ \textit{s.t } \psi_i \cap \psi_{i_t} = \emptyset\}$$

Then the **degree of coherence** *of K is defined as follows where $A(S) = \Sigma_{(\psi_i, \kappa_i) \in S} \ \kappa_i$*

$$\mathsf{Coherence}(K) = 1 - \frac{A(\mathsf{ConflictBase}(K))}{A(\mathsf{OpinionBase}(K))}$$

Proposition 4. *Let K be a PVC. If the possibility distribution associated with this PVC is normal, then* $\mathsf{Coherence}(K) = 1$.

When a PVC produces a normal possibility distribution, the weighted subsets in the PVC share at least one common element, therefore, the ConflictBase is empty which results in a degree of coherence of 1.

Proposition 5. *Let K be a PVC. If K is inconsistent with low quality, then* $\mathsf{Coherence}(K) = 0$.

When a PVC is inconsistent with lower quality, every weighted subset in the PVC is selected in the ConflictBase, which is in turn equal to the OpinionBase, and therefore, the degree of coherence is 0.

Now, we use this new measure to examine the two PVCs in Example 2. Let K_1 and K_2 denote the two PVCs left and right respectively, the coherence measures of the two PVCs are

$\mathsf{Coherence}(K_1) = 1 - \frac{\Sigma_{i=1}^{p} \kappa_i}{\Sigma_{i=1}^{p} \kappa_i} = 0; \quad \mathsf{Coherence}(K_2) = 1 - \frac{\Sigma_{i=1,2,4} \kappa_i}{\Sigma_{i=1}^{4} \kappa_i} = 3/11$

It is obvious that although the two PVCs have the same degree of inconsistency (e.g., 0.3), they have different degrees of coherence measure. The quality of K_2 is better than that of K_1 because the subsets that are assigned with degrees of belief (in terms of necessity measures) in K_2 are largely overlap whilst the subsets with degrees of belief in K_1 are distinct which suggests that this knowledge is more contradicting internally.

The above defined coherence measure includes a weighted subset (e.g., (ψ_i, κ_i)) in the ConflictBase as long as there exists another weighted subset that the intersection of them is empty, although ψ_i may share some common elements with all other subsets. Obviously, there can be many ways to define a conflict base, and the one defined in Definition 7 above is the largest in terms of cardinality.

On the other hand, the smallest conflict base possible is to include those weighted subsets which have no intersection with any other weighted subsets. This will surely result in a higher degree of coherence comparing to a larger conflictbase. Below, We give the definition of this conflict base and its corresponding coherence measure and call this measure the upper bound of the degree of coherence.

Definition 8. *Let K be a PVC.*

$$\mathsf{OpinionBase}(K) = \{(\psi_i, \kappa_i) | \textit{ such that } (\psi_i, \kappa_i) \textit{ is a weighted subset of } K \}$$

$$\mathsf{UpperConflictBase}(K) = \{(\psi_i, \kappa_i) \in \mathsf{OpinionBase}(K) | \forall (\psi_{i_t}, \kappa_{i_t}) \in \mathsf{OpinionBase}(K) \textit{ if } \psi_i \neq \psi_{i_t} \textit{ then } \psi_i \cap \psi_{i_t} = \emptyset\}$$

Then the **upper bound of the degree of coherence** *of K is defined as follows where $A(S) = \Sigma_{(\psi_i, \kappa_i) \in S} \ \kappa_i$.*

$$\mathsf{UpperCoherence}(K) = 1 - \frac{A(\mathsf{UpperConflictBase}(K))}{A(\mathsf{OpinionBase}(K))}$$

It is easy to verify that Propositions 5 and 6 are still valid with $\mathsf{UpperCoherence}(K)$, since $\mathsf{UpperCoherence}(K)$ is always greater than $\mathsf{Coherence}(K)$. Interval

$$[\mathsf{Coherence}(K), \mathsf{UpperCoherence}(K)]$$

of a PVC defines the range of its coherence measure with following properties.

- when $[\mathsf{Coherence}(K), \mathsf{UpperCoherence}(K)] = [1, 1]$, the PVC is totally coherent. For example, when the associated possibility distribution of a PVC is normal, the corresponding coherence measure interval is $[1, 1]$. However, a $[1, 1]$ interval does not guarantee a PVC having a normal possiblity distribution. For instance, a PVC with three weighted subsets $\{(\{1, 2, 4\}, 0.5), \{2, 3\}, 0.4), \{3, 4\}, 0.7)\}$ has interval $[1, 1]$, but its possibility distribution is not normal (where numerical numbers are the indexes for elements in the associated frame).
- when $[\mathsf{Coherence}(K), \mathsf{UpperCoherence}(K)] = [0, 0]$, the PVC is inconsistent with lower quality, see Proposition 6.
- when $[\mathsf{Coherence}(K), \mathsf{UpperCoherence}(K)] = [\alpha, 1]$ where $\alpha > 0$, the PVC has some weighted subsets that is not in conflict with any other subsets. An example is when a PVC is inconsistent with good quality and has no isolated separable elements. The right PVC in Example 5 specifies this case with the interval $[3/11, 1]$.
- when $[\mathsf{Coherence}(K), \mathsf{UpperCoherence}(K)] = [0, b < 1]$, the PVC has at least one isolated separable element.
- when $[\mathsf{Coherence}(K), \mathsf{UpperCoherence}(K)] = [0, 1]$. Any other situations not falling into the above categories.

For the last case where the pair gives $[0, 1]$ interval, there can be many situations to provoke this situation as illustrated by the next example.

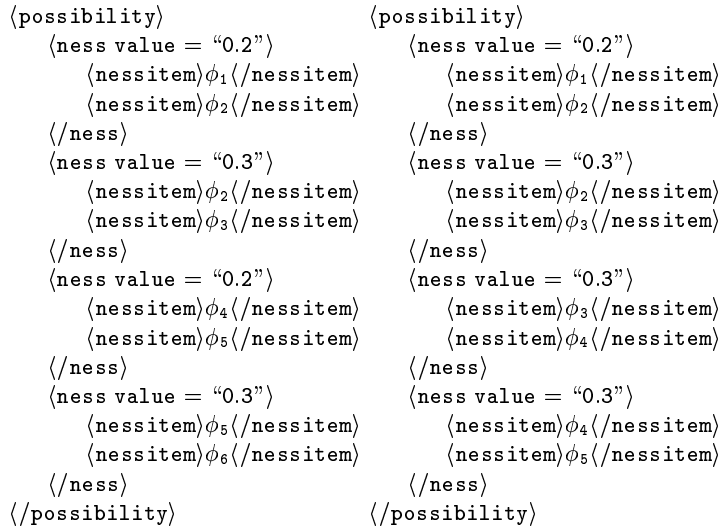

Fig. 4. Two possibility-valid components (PVCs) ($\Omega = \{\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6\}$).

Example 3. Consider Figure 4. Both of the PVCs have the same degree of inconsistency and the same interval of the degrees of coherence. The left PVC forms two separate clusters, whilst the right PVC forms a chain of subsets with each neighbouring pair sharing one comment element. At present, our methods for measuring coherence cannot distinguish the quality between these two situations.

Coherence measures are useful additions to the concept of degree of inconsisitency, since they provide more information about the quality of an XML document when a degree of inconsistency is not sufficient. These measures can be used to rank information from multiple sources when no extra data is available about their reliablity.

Definition 9. *Let $\leq$ on the set $\{[1,1], [\alpha,1], [0,1], [0,\beta], [0,0]\}$ (where $1 > \alpha > 0$ and $1 > \beta > 0$) be a binary relation such that*

$[0,0] \leq [0,\beta]$*;* $[0,\beta] \leq [0,1]$*;* $[0,1] \leq [\alpha,1]$*;* $[\alpha,1] \leq [1,1]$*;*
$[\alpha_1,1] \leq [\alpha_2,1]$ *if* $\alpha_1 \leq \alpha_2$*;*
$[0,\beta_1] \leq [0,\beta_2]$ *if* $\beta_1 \leq \beta_2$*.*

$\leq$ *is a lex-ordering.*

Proposition 6. *Let K be a PVC with coherence interval $[\alpha,\beta]$. When $\alpha > 0$, $\beta = 1$ and when $\beta < 1$, $\alpha = 0$.*

Proof: *When* $\mathsf{Coherence}(K) = \alpha > 0$ *is true, it implies that there exists at least one weighted subset, (ψ_i, κ_i), such that for any other weighted subset (ψ_j, κ_j), $\psi_i \cap \psi_j \neq \emptyset$, and ψ_i is not included in the* $\mathsf{ConflictBase}(K)$*. It further implies that there is no isolated*

separable element in this component, otherwise, the intersection of ψ_i with this isolated separable element would have been empty. Therefore, $\mathsf{UpperConflictBase}(K) = \emptyset$*, and* $\mathsf{UpperCoherence}(K) = \beta = 1$.

On the other hand, when $\beta < 1$ it implies that there is at least on isolated separable element, such that it has no common element with any other weighted subset. Therefore, every weighted subset is selected in $\mathsf{ConflictBase}(K)$*, and $\alpha = 0$.* □

With this proposition, together with the fact that the $\leq$ relation is a partial order relation, we see that Definition 9 is sufficient to cover all the possible intervals of coherence measures of PVCs.

Definition 10. *Let K_1 and K_2 be two PVCs with the same degree of inconsistency. Let I_{K_1} and I_{K_2} be two elements in the set in Definition 9 representing the intervals of coherence measures of K_1 and K_2 respectively. PVC K_2 is said to be more coherent than K_1 if $I_{K_1} \leq I_{K_2}$.*

Based on this partial order relation on X, it is possible to rank any number of information sources by ranking the quality of their PVCs.

4 Conclusion

In this paper, we have proposed some definitions and a coherence based method to assess the quality of an inconsistent PVC when the degree of inconsistency alone is not adequate to serve the purpose. The coherence based method can be used to rank information sources based on the quality of the information they provide. A potential application of the method is in information fusion where multiple PVCs need to be merged. When no preferences are given about information sources, information from highly ranked PVC could be merged before that of lower ranked ones if the sequence of merging is of an importance. Furthermore, the coherence measures can be used to select a more appropriate merging operator to merge a set of PVCs. For instance, given four PVCs which are pair-wise inconsistent, a disjunctive operator, e.g., max, is usually used to merge them which may result in an almost uniform possibility distribution. The merged result provides less information than the original sources. However, if the coherence measures of the conjunctively merged PVC suggest that the PVC is largely coherent, e.g., with a coherent interval $[\beta, 1]$, then applying the conjunctive operator may be of a better choice than the disjunctive one. The preliminary result of our investigate into this topic is summarized in [HL05c].

The measures of quality may also be used to assess whether a PVC should be rejected prior to merging. For example, suppose we have a set of news reports to merge, and suppose each news report is represented by a structured report, and further suppose each strucutured report contains a PVC with key information, then we may choose to ignore the structured reports with PVCs of low quality, or may send them back to their supplier with a request for clarification.

The two definitons on judging whether a PVC is of a *good* or *lower* quality, given that it is inconsistent, provides a way of assessing its quality without calculating its coherence intervals. A useful extension of the definition on *good* quality PVC is the new normalization rule that is best suited for this situation.

Our definitions of coherence measures can be seen as extensions of the coherence function in [DKP03] where this function is defined in a Quasi-possibilistic logic framework. The definitions of the ConflictBase and the OpinionBase are based on the quasi-classical interpretations of the given knowlege base. We inherited the spirit of the function, but provided new definitions of the ConflictBase and the OpinionBase, as well as the UpperConflictbase in set based situations.

Less closely related work is that on measuring the impression of a possibility distributtion π ([DP87b], [HK83]), denoted as $\mathsf{Imp}(\pi)$. This measure was defined only when the possibilistic knowledge base associated with π was consistent. For an inconsistent situation, $\mathsf{Imp}(\pi)$ was recalculated as $\mathsf{Imp}(\pi)/(1 - \mathsf{Inc}(K))$.

References

[BDP97] S Benferhat, D Dubois, and H Prade. From semantic to syntactic approach to information combination in possibilitic logic. In *Aggregation and Fusion of Imperfect Information*, pages 141-151. Physica Verlag, 1997.

[BDKP00] S Benferhat, D Dubois, S Kaci, and H Prade. Encoding classical fusion in ordered knowledge bases framework. In *Linkping Electronic Articles in Computer and Information Science*, Vol. 5, No. 027, 2000.

[DKP03] D Dubois, S Konieczny and H Prade. Quasi-possibilistic logic and its measures of information and conflict. *Fundamenta Informaticae*, Vol. 57:101-125, 2003.

[DP87a] D Dubois and H Prade. The principle of minimum specificity as a basis for evidential reasoning. *Uncertainty in Knowledge-Based Systems*, Bouchon and Yager (Eds.). Springer-Verlag, pages 75-84, 1987.

[DP87b] D Dubois and H Prade. Properties of measures of information in evidence and possibility theories. *Fuzzy Sets and Systems*, Vol. 24:161-182,1987.

[DP88] D Dubois and H Prade. *Possibility theory: An approach to the computerized processing of uncertainty*. Plenum Press, 1988.

[HK83] M Higashi and G Klir. Measures of uncertainty and information based on possibility distributions. *International Journal of General Systems*, Vol. 9: 43-58, 1983.

[HL04] A Hunter and W Liu. Logical reasoning with multiple granularities of uncertainty in semi-structured information. *Procedings of IPMU'04*, 1009-1016. 2004.

[HL05a] A Hunter and W Liu. Fusion rules for merging uncertain information. *Information Fusion Journal*. (in press) 2005

[HL05b] A Hunter and W Liu. Merging uncertain information with semantic heterogeneity in XML. *Knowledge and Information Systems*. (to appear) 2005.

[HL05c] A Hunter and W Liu. Assessing the quality of merged information in possibilistic XML. Technical Report, Department of Computer Science, UCL. 2005.

[HS04] A Hunter and R Summerton. Fusion rules for context-dependent aggregation of structured news reports. *Journal of Applied Non-classical Logic*, 14(3):329–366, 2004.

[Hun02] A Hunter. Measuring inconsistency in knowledge via quasi-classical models. *Proceedings of AAAI'2002*, 68–73, 2002.

[KKA05] M van Keulen, A de Keijzer and W Alink A probabilistic XML approach to data integration. *Proceedings of ICDE'05*, 2005.

[NJ02] A Nierman and H Jagadish. ProTDB: Probabilistic data in XML. In *Proceedings of VLDB'02, LNCS* 2590, pages 646–657. Springer, 2002.

[SDK95] S Sandri, D Dubois, and H Kalfsbeek. Elicitation, assessment and polling of expert judgements using possibility theory. *IEEE Trans on Fuzzy Systems*, 3:313–335, 1995.