A logical reasoning framework for modelling and merging uncertain semi-structured information

Anthony Hunter [a] **and Weiru Liu** [b]

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

Abstract

Semi-structured information in XML can be merged in a logic-based framework [7,9]. This framework has been extended to deal with uncertainty, in the form of probability values, degrees of beliefs, or necessity measures, in the XML documents [8]. In this paper, we discuss how this logical framework can be used to model and reason with structured scientific knowledge on the Web in medical and bioscience domains. We will demonstrate how multiple summaritive and evaluative knowledge under uncertainty can be merged to obtain less conflicting and better confirmed results in response to users queries. We will also show how reliability of a source can be integrated into this structure.

Key words: Semi-structured information fusion, uncertain information in XML

1. Introduction

XML has been used extensively on the Web for representing and exchanging a variety of static and dynamic information, such as database query results. Along with its increasing use in a wider range of activities, the need to represent uncertain information has rapidly emerged recently, since in real life, information is often uncertain and incomplete.

Two typical examples of integrating uncertainty into the XML structure are [6] and [10], both methods assign probabilistic values to elements in an XML document. A probability value can either be assigned to a leaf node (a textentry) or a tagname, but these two approaches offer different methods to calculate a final probability for a query of XML information. Another attempt to model uncertainty in XML is reported in [1] where numerical values representing the importance of tags are attached to tagnames. These values are interpreted in fuzzy theory and used to calculate the importance of a set of tagnames in comparison to other sets of tagnames, so that more important information can be used first to make decisions. Its primary application domain is service related information gathering in which customers have choices over a set of options. With a decision tailored to options that are more important, a customer is more likely to be satisfied by the service provided. Also, under the umbrella of making Web information more meaningful, a proposal was reported in [11] which integrates probabilities into DAML+OIL, a commonly used ontology language in the Semantic Web. Uncertain statements are marked with probability values instead of assuming that every statement is either true or false as in the current language format.

In contrast to the approaches above, our logic-based framework aims at establishing a formal structure that can facilitate uncertainty reasoning in formal logics that in turn make use of knowledge in the background knowledgebase to assist querying and merging. The framework has proved to be capable of modelling a variety of forms of uncertainty and has advantages over both approaches [6,10].

In this paper, we discuss how this extended logical framework can be used for modelling and reasoning with structured scientific knowledge on the Web in medical and bioscience domains. We will demonstrate how multiple summaritive and evaluative knowledge under uncertainty can be easily merged to obtain less conflicting and better confirmed results in response to users queries. A number of examples are deployed to illustrate potential applications of the framework. We will proceed as follows. Section 2 discusses what constitutes structured scientific knowledge and the need for modelling uncertainty in XML. Section 3 reviews the basic definitions in the logical fusion framework with examples. Section 4 investigates how the reliability of a source can be explicitly represented in the framework and how it is integrated with other types of uncertainty in the process of answering a user's query. Section 5 looks at the issue of merging multiple XML documents for both probability values and mass functions. Section 6 summarizes the paper.

2. Structured scientific knowledge

Structured scientific knowledge We use XML documents to represent semi-structured information such as structured scientific knowledge (SSK). Each SSK report describes information in one or more scientific datasources (such as journals, databases of empirical results, etc). The format of an SSK report is an XML document. Each SSK report contains **summaritive information** about the datasource

(e.g. information from an abstract, summary of techniques used, etc) plus **evaluative information** about the datasource (eg. delineation of uncertainties and errors in the information source, qualifications of the key findings, etc). Each SSK report can be constructed by hand, by information extraction systems (e.g. [3]), or as the result of querying and analysing scientific databases in [10,12].

For instance, in the medical community, the number of journals and conferences having articles that are relevant to a single specific topic is extremely large and fast growing. This makes it very difficult for physicians to keep pace of all new results reported in their fields and makes it even harder for patients to find relevant information. There is an increasing need to better summarize this raw information so that different types of user can get more satisfactory summaritive and evaluative information.

In [5], a system called *Persival* was developed which aims at providing tailored presentation of relevant medical literature for both physicians and lay consumers. Based on a user's query, the system takes documents (including images and video) as input, and generates one or more paragraphs of summary from the input documents, highlighting the common points and the differences among these input documents. The summaries can also be provided at different levels of granularity depending on who the user is. Each summary follows a fixed structure including *introduction, methods, results, and discussion.* For documents with patient medical records, the output is in a more structured format which can be easily represented with XML documents. Already in [12], the query results of medical journals are directly expressed as XML documents and these results are merged to reduce incompleteness and error messages.

Another source of SSK reports can be obtained as a by-product of querying databases. There are many online information resources for bioinformatics. Most of the information in these sites is in a semi-structured format. For example, when searching for information related to a specific protein, a bioscientist may invoke specialised database search tools, such as BLAST. The results of such searches are in semi-structured format and may need to be saved by the user and then searched. It is desirable to collect these results and extract summary information from them. Such information may then be integrated with searches of abstracts such as those stored in PubMed.

As more and more individual SSK reports accumulate, there is an urgent need to integrate them. An example of integrating query results in XML format is reported in [13] where the main focus is on semantic integration of life science databases. As it was argued that publically available biological knowledge is scattered over many hundred internet accessible data sources, data integration is a fundamental prerequisite for answering complexity queries. Another example of this kind is from [2] which focuses on merging temporal aspects of multimedia semi-structured data in clinical information. Temporal clinical semi-structured information is first modelled in a graphical model and then translated into XML documents.

In summary, with XML being increasingly used as a standard data exchange format, integrating information in XML documents is a pressing task in making the best use of available data sources. Furthermore, in real-world applications, many

summaritive and evaluative information and query results are often subject to uncertainty and inconsistency. Therefore, an automated XML integration tool should be able to deal with inconsistencies and uncertainties in information when they arise.

In repsonse to this need, we have been developing a logic-based fusion framework that supports context-dependent representation and reasoning involving uncertainty in information. In our approach, each SSK report is regarded as a tree and this can isomorphically be represented as a logical term. Therefore, logical reasoning technologies can be applied. A query of merging some SSK reports can be handled by recursive calls to a logical reasoning tool to merge the subtrees in the SSK reports. This gives a context-dependent logic-based approach to merging that is sensitive to the uncertain information in the SSK reports and to the background knowledge in the knowledgebase. The apparent structural difference of multiple XML documents can be resolved by using XSLT which is able to transform one XML document into the format of another XML document. Therefore, in this paper, we don't consider structural heterogeneity in XML documents.

Uncertainty in XML An important feature of SSK reports is the ability to represent **uncertainty**. Much leading-edge scientific information is subject to uncertainty, and of diverse types, including empirical methods (such as the nature of populations and samples, estimates of experimental errors, etc), statistical analysis (such as mean, standard deviation, probability statements, correlation, significance tests, etc), and subjective assessments drawn on the basis of the evidence. For example, a probability distribution over a set of possible outcomes as textentries $\tau_1, .., \tau_n$ for a tagname ϕ, where $\mathtt{x_i}$ is the probability of $\tau_\mathtt{i}$, can be represented by the following piece of XML that would be nested in an SSK report.

$$\langle\phi\rangle\langle\mathtt{prob\ value} = \text{``}\mathtt{x_1}\text{''}\rangle\tau_1\langle/\mathtt{prob}\rangle...\langle\mathtt{prob\ value} = \text{``}\mathtt{x_n}\text{''}\rangle\tau_\mathtt{n}\langle/\mathtt{prob}\rangle\langle/\phi\rangle$$

Encoding probabilities (or uncertainty values) involves refining the DTD for SSK reports to enforce the use of specific tags for uncertain information. We represent uncertainty in the XML for SSK reports as developed in [8] where uncertainty can be modelled in either probability theory, belief function theory [14], or possibility theory [4].

The primary objective of **merging** SSK reports is to decrease redundancy between SSK reports, to address incompleteness in individual SSK reports, and most importantly to minimize the inconsistencies and uncertainties arising in SSK reports. So that merging would provide a better and more complete summary and evaluation of the datasources involved. For example, if two SSK reports are on the same subject and they are mutually conflicting, i.e the union of them is highly inconsistent, then they reveal that either one or both sources are not correct. This can then be a qualification assigned to the evaluative information in the SSK reports that indicates there is a problem with one or both datasources. This can be very useful especially for empirical data in datasources.

3. A logical fusion framework

We review some of the basic definitions in the framework [8].

Definition 1 Structured Scientific Knowledge Report (SSK report): *If φ is a tagname (i.e an element name), and ϕ is textentry, then $\langle\varphi\rangle\phi\langle/\varphi\rangle$ is an SSK report. If φ is a tagname, 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 an SSK report. If φ is a tagname and $\sigma_1, ..., \sigma_n$ are SSK reports, then $\langle\varphi\rangle\sigma_1...\sigma_n\langle/\varphi\rangle$ is an SSK report.*

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

Clearly each SSK report is isomorphic to a tree with the non-leaf nodes being the tagnames and the leaf nodes being the textentries. This isomorphism allows us to give a definition for an *abstract term* of an SSK report. Via this isomorphic relationship, we can refer to a branch of an abstract term by using the branch of the isomorphic SSK, and we can refer to a subtree of an abstract term by using the subtree of the isomorphic SSK.

Definition 1 describes how an XML document can be defined recursively starting from the simplest one which has only one tagname and one value associated with the tagname. Definition 2 defines how a tree structure like XML document can be equally described as a logical term which also reflects the relationships between tagnames and their values. For instance, XML information $\langle\texttt{date}\rangle\texttt{03/03/99}\langle/\texttt{date}\rangle$ is denoted as $\texttt{date(03/03/99)}$ in logics where $\texttt{03/03/99}$ can be understood as the value of attribute $\texttt{date}$.

We consider two types of uncertainty in this paper, probability values and mass functions in DS theory [14]. A mass function, m, is defined on a set of mutually exclusive and exhaustive set of values Ω called a *frame of discernment* (or simply *frame*), as $m(\emptyset) = 0$ and $\Sigma_{A\subseteq\Omega}\ m(A) = 1$. The formal modelling approach to representing these two types of uncertainty is given in the following two definitions.

Definition 3 *The SSK report $\langle\texttt{probability}\rangle\sigma_1,..,\sigma_n\langle/\texttt{probability}\rangle$ is a* **probability-valid component (ProVC)** *iff each $\sigma_i \in \{\sigma_1,..,\sigma_n\}$ is of the form $\langle\texttt{prob value} = \kappa\rangle\ \phi\langle/\texttt{prob}\rangle$ where $\kappa \in [0,1]$ and ϕ is a textentry.*

Definition 4 *The SSK report $\langle\texttt{belfunction}\rangle\ \sigma_1,..,\ \sigma_n\langle/\texttt{belfunction}\rangle$ is a* **belfunction-valid component (BelVC)** *iff for each $\sigma_i \in \{\sigma_1,..,\sigma_n\}$ σ_i is of the form $\langle\texttt{mass value} = \kappa\rangle\sigma_1^i,...,\sigma_m^i\langle/\texttt{mass}\rangle$ and for each $\sigma_j^i \in \{\sigma_1^i,..,\sigma_m^i\}$, σ_j^i is of the form $\langle\texttt{massitem}\rangle\phi\langle/\texttt{massitem}\rangle$ where $\kappa \in [0,1]$ and ϕ is a textentry.*

```
⟨report⟩
  ⟨prostate cancer prediction⟩
  ⟨reliability = "0.7"⟩
    ⟨author⟩unknown ⟨/author⟩
    ⟨title⟩Prostatic Specific Antigen Screening Test⟨/title⟩
    ⟨url⟩http : //medic.med.uth.tmc.edu/ptnt/00000390.htm⟨/url⟩
      ⟨PSA range = "0.0 − 3.9"⟩
        ⟨conclusion⟩NoCancer⟨/conclusion⟩
      ⟨/PSA⟩
      ⟨PSA range = "4.0 − 9.9"⟩
        ⟨conclusion⟩
          ⟨probability⟩
            ⟨prob value = "0.22"⟩Cancer⟨/prob⟩
            ⟨prob value = "0.78"⟩NoCancer⟨/prob⟩
          ⟨/probability⟩
        ⟨/conclusion⟩
      ⟨/PSA⟩
      ⟨PSA range > "10.0"⟩
        ⟨conclusion⟩
          ⟨probability⟩
            ⟨prob value = "0.65"⟩Cancer⟨/prob⟩
            ⟨prob value = "0.35"⟩NoCancer⟨/prob⟩
          ⟨/probability⟩
        ⟨/conclusion⟩
      ⟨/PSA⟩
  ⟨/reliability⟩
  ⟨/prostate cancer prediction⟩
⟨/report⟩
```

Fig. 1. An XML document with uncertain information

All textentries in the above two definitions are elements of a pre-defined set Ω in the background knowledgebase. We also require that $\Sigma_i \kappa_i = 1$ for both cases to preserve the constrains in both theories.

Let us take prostate cancer prediction and diagnosis as an example. There are two types of methods available for users to get some initial information. One method is based on the Prostate Specific Antigen (PSA) value through a blood test. Higher PSA values can flag the possibility of cancer. However, this method is subject to inaccuracy, due to the fact that a higher PSA value can be influenced by many other factors, such as prostate inflammation and horse riding, before taking the blood sample. In general, this method is about 70% accurate in cancer diagnosis (http://medic.med.uth.tmc.edu/ptnt/00000390.htm). This high level summary can be represented in an XML document as shown in Figure 1.

In this example, we use two ProVCs to represent the conclusions drawn form certain PSA values. Furthermore, since this diagnosis is not absolutely accurate, we insert a reliability factor into the XML to indicate how much credence we should give to this piece of information. Obviously, there is a need to formalize the reliability factor into the diagnostic result. We look at this issue next.

4. Integrating reliability of SSK into XML

The reliability value in the above example is different from probability distributions on text entries such as Cancer or NoCancer. The former identifies how reliable a conclusion, or a source, or an experiment, is. In this section, we investigate the method to integrate this factor with other uncertainty components (ProVCs, BelVCs) in XML documents when answering a query For this purpose, we look at the discounting operator in DS theory.

Discounting is useful and essential when a belief function (or a mass function) fails to take into account some particular uncertainty affecting the evidence as a whole [14]. Assume that the evidence is accurate to α degree, then the information provided by the evidence should be discounted by degree $1 - \alpha$. Let m be a mass function on Ω provided by a piece of evidence which is in turn has the degree of reliability (or trust) α, then a new mass function m' defined by

$$m'(A) = \begin{cases} \alpha m(A) & \text{when } m(A) > 0, \ A \subset \Omega \\ (1 - \alpha) + \alpha m(A) & \text{when } A = \Omega \end{cases}$$

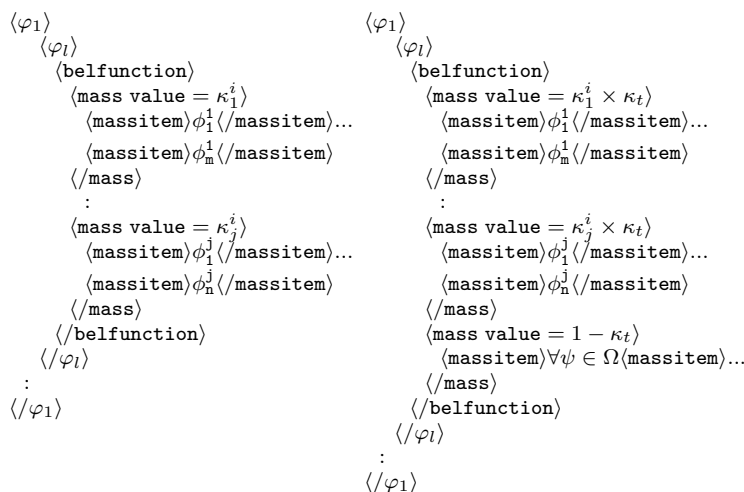has taken into account the impact of imprecision of the evidence.

Definition 5 *A SSK report* $\langle \sigma_1 \rangle \langle \mathtt{reliability} = \kappa^1 \rangle \sigma_1^1, .., \sigma_n^1 \langle / \mathtt{reliability} \rangle \langle / \sigma_1 \rangle$ $, .., \langle \sigma_t \rangle \langle \mathtt{reliability} = \kappa^t \rangle \sigma_1^t, .., \sigma_m^t \langle / \mathtt{reliability} \rangle \langle / \sigma_t \rangle$ *is a* **reliability-valid component (RelVC)** *where* $\kappa^i > 0$ *and each* $\sigma_i^l \in \{\sigma_1^1, \ldots, \sigma_n^1\} \cup \ldots \cup \{\sigma_1^t, \ldots, \sigma_m^t\}$ *is a valid SSK report.*

Definition 6 *Let a section of a RelVC for tag* σ *be* $\langle \sigma \rangle \langle \mathtt{reliability} = \kappa_t \rangle \sigma_1, .., \sigma_n$ $\langle / \mathtt{reliability} \rangle \langle / \sigma \rangle$ *and any* σ_i *does not contain any further RelVC components. Let* σ_i *be a BelVC with structure in Figure 1 left, then the transformed* σ_i' *defined in Figure 2 right is a BelVC incorporating the value of the reliability. The original section of the RelVC for tag* σ *is thus revised as* $\langle \sigma \rangle \sigma_1', .., \sigma_n' \langle \sigma \rangle$, *where each* σ_i' *has the reliability factor being integrated.*

The last section with $\langle \mathtt{mass \ value} = 1 - \kappa_t \rangle$ is the value assigned to the frame Ω, if there is no mass value assigned to it in the initial XML document. Otherwise, this section has appeared as a σ_j above and should not be added again here.

Starting with an XML document that contains both reliability factors and uncertainty components, the above definition generates a new XML document from it consisting of only uncertainty components,

Example Since a PSA value only provides an approximate prediction and suffers from drawbacks of inaccuracy, more comprehensive methods have been proposed to analyze patient's tests results thoroughly. Here, we look at one of such methods. Assume that for each patient, there is a blood serum mass spectrum. The features of a spectrum are defined as the x-axis locations within the spectrum that are able to distinguish healthy and cancerous status based on y-axis values. The full resolution of a spectrum can contain 15000 features and may be subject to noise. Usually, it is possible to smooth the spectrum to produce less features and to reduce noise. Commonly, feature numbers are reduced by half in each smoothing stage.

$\langle\varphi_1\rangle$
$\quad\langle\varphi_l\rangle$
$\qquad\langle\text{belfunction}\rangle$
$\qquad\quad\langle\text{mass value} = \kappa_1^i\rangle$
$\qquad\qquad\langle\text{massitem}\rangle\phi_1^1\langle/\text{massitem}\rangle...$
$\qquad\qquad\langle\text{massitem}\rangle\phi_m^1\langle/\text{massitem}\rangle$
$\qquad\quad\langle/\text{mass}\rangle$
$\qquad\qquad\vdots$
$\qquad\quad\langle\text{mass value} = \kappa_j^i\rangle$
$\qquad\qquad\langle\text{massitem}\rangle\phi_1^j\langle/\text{massitem}\rangle...$
$\qquad\qquad\langle\text{massitem}\rangle\phi_n^j\langle/\text{massitem}\rangle$
$\qquad\quad\langle/\text{mass}\rangle$
$\qquad\langle/\text{belfunction}\rangle$
$\quad\langle/\varphi_l\rangle$
$\quad\vdots$
$\langle/\varphi_1\rangle$

$\langle\varphi_1\rangle$
$\quad\langle\varphi_l\rangle$
$\qquad\langle\text{belfunction}\rangle$
$\qquad\quad\langle\text{mass value} = \kappa_1^i \times \kappa_t\rangle$
$\qquad\qquad\langle\text{massitem}\rangle\phi_1^1\langle/\text{massitem}\rangle...$
$\qquad\qquad\langle\text{massitem}\rangle\phi_m^1\langle/\text{massitem}\rangle$
$\qquad\quad\langle/\text{mass}\rangle$
$\qquad\qquad\vdots$
$\qquad\quad\langle\text{mass value} = \kappa_j^i \times \kappa_t\rangle$
$\qquad\qquad\langle\text{massitem}\rangle\phi_1^j\langle/\text{massitem}\rangle...$
$\qquad\qquad\langle\text{massitem}\rangle\phi_n^j\langle/\text{massitem}\rangle$
$\qquad\quad\langle/\text{mass}\rangle$
$\qquad\quad\langle\text{mass value} = 1 - \kappa_t\rangle$
$\qquad\qquad\langle\text{massitem}\rangle\forall\psi \in \Omega\langle\text{massitem}\rangle...$
$\qquad\quad\langle/\text{mass}\rangle$
$\qquad\langle/\text{belfunction}\rangle$
$\quad\langle/\varphi_l\rangle$
$\quad\vdots$
$\langle/\varphi_1\rangle$

Fig. 2. Transformation of a reliability factor

For the sample data at http://clinicalproteomics.steem.com/download-prost.php, a Bayesian Classifier program (by Dr. Cheng) runs these data first with a full resolution giving 15000 features, then with lower resolutions having $15000/2$ (giving 7500 features), $15000/2^2$, $15000/2^3$, and $15000/2^4$ features respectively. The relative accuracies of cancer diagnosis under these five resolutions are 90%, 92.68%, 91.75%, 92.33%, and 85.81% respectively. The experimental result of this analysis is then summarized in the following XML document.

```
⟨report⟩
  ⟨prostate cancer prediction⟩
    ⟨author⟩Jie Cheng⟨/author⟩
    ⟨title⟩Bayesian Classifier of prostate cancer⟨/title⟩
    ⟨url⟩http : //clinicalproteomics.steem.com/download − prost.php⟨/url⟩
    ⟨dataName⟩accuracy of blood serum mass spectrum⟨dataName⟩
      ⟨features = 15000⟩
       ⟨conclusionAccuracy⟩ "0.9" ⟨/conclusionAccuracy⟩
      ⟨/features⟩
      ⟨features = 7500⟩
       ⟨conclusionAccuracy⟩ "0.9268" ⟨/conclusionAccuracy⟩
      ⟨/features⟩
      ⟨features = 3750⟩
       ⟨conclusionAccuracy⟩ "0.9175" ⟨/conclusionAccuracy⟩
      ⟨/features⟩
      ⟨features = 1875⟩
       ⟨conclusionAccuracy⟩ "0.9233" ⟨/conclusionAccuracy⟩
      ⟨/features⟩
      ⟨features = 937⟩
       ⟨conclusionAccuracy⟩ "0.8581" ⟨/conclusionAccuracy⟩
      ⟨/features⟩
  ⟨/prostate cancer prediction⟩
⟨/report⟩
```

When a patient's spectrum is analysed using this classifier under a specific resolution, for instance, the full resolution, a conclusion will be drawn as to whether the patient has cancer. Assume that the conclusion is Cancer, then the degree of accuracy of this analysis under *this* resolution shall be taken into account, so the conclusion is revised as *the patient is having cancer with chance* 90%. This statement also implies that with 10% chance we do not know what the conclusion would be, e.g., either Cancer or NoCancer.

This XML document can be used to derive diagnosis for individual patients. For example, assume that patient J Sky's spectrum is known and features have been selected under the full resolution. Feeding these values into the Bayesian Classifier, a diagnosis will be conducted with a probability attached to each of the two possible outcomes, Cancer or NoCancer. The corresponding XML document is as follows.

```
⟨report⟩
  ⟨prostate cancer prediction⟩
   ⟨author⟩Jie Cheng⟨/author⟩
   ⟨title⟩Bayesian Classifier ⟨/title⟩
   ⟨patient⟩J. Sky⟨/patient⟩
   ⟨date⟩06/11/2003⟨/date⟩
   ⟨dataName⟩blood serum mass spectrum⟨dataName⟩
    ⟨features = 15000⟩
     ⟨reliability = "0.9"⟩
        ⟨conclusion⟩
          ⟨probability⟩
            ⟨prob value = "0.4985569"⟩Cancer⟨/prob⟩
            ⟨prob value = "0.5014431"⟩NoCancer⟨/prob⟩
          ⟨/probability⟩
        ⟨/conclusion⟩
     ⟨/reliability⟩
    ⟨/features⟩
  ⟨/prostate cancer prediction⟩
⟨/report⟩
```

Since a ProVC can be seen as a special case of BelVC and we have a predicate to convert a ProVC into a BelVC [8], it is possible to first convert the ProVC for J Sky into a BelVC and then apply Definition 6 to generate an XML document with the reliability degree being integrated into the BelVC . The newly derived BelVC gives $m(\mathsf{Cancer}) = 0.44870121, m(\mathsf{NoCancer}) = 0.45129879, m(\mathsf{Cancer}, \mathsf{NoCancer}) = 0.1$ and the BelVC segment is

```
⟨belfunction⟩
  ⟨mass value = "0.44870121"⟩
     ⟨massitem⟩Cancer⟨/massitem⟩
  ⟨/mass⟩
  ⟨mass value = "0.45129879"⟩
     ⟨massitem⟩NoCancer⟨/massitem⟩
  ⟨/mass⟩
  ⟨mass value = "0.1"⟩
     ⟨massitem⟩Cancer⟨/massitem⟩
     ⟨massitem⟩NoCancer⟨/massitem⟩
  ⟨/mass⟩
⟨/belfunction⟩
```

5. Merging multiple uncertainty information

Merging occurs when multiple sources of information available concerning the same issue. We first review the predicate for merging two BelVCs.

Definition 7 *([8]) Let* $\langle\texttt{belfunction}\rangle\sigma_1^1,..,\sigma_p^1\langle/\texttt{belfunction}\rangle$ *and* $\langle\texttt{belfunction}\rangle$ $\sigma_1^2,..,\sigma_q^2\langle/\texttt{belfunction}\rangle$ *be two BelVCs, where*

   (i) $\sigma_i^1 \in \{\sigma_1^1,..,\sigma_p^1\}$ *is of the form* $\langle\texttt{mass value} = \kappa_i^1\rangle\psi_i^1\langle/\texttt{mass}\rangle$

   (ii) the (subset, mass) pair collection is $S_1 = \{(\psi_1^1, \kappa_1^1), \ldots, (\psi_p^1, \kappa_p^1)\}$,

   (iii) $\sigma_j^2 \in \{\sigma_1^2,..,\sigma_q^2\}$ *is of the form* $\langle\texttt{mass value} = \kappa_j^2\rangle\psi_j^2\langle/\texttt{mass}\rangle$

   (iv) the (subset, mass) pair collection is $S_2 = \{(\psi_1^2, \kappa_1^2), \ldots, (\psi_q^2, \kappa_q^2)\}$,

Let the **combined BelVC** *be* $\langle\texttt{belfunction}\rangle\sigma_1,..,\sigma_s\langle/\texttt{belfunction}\rangle$ *where each* $\sigma_k \in \{\sigma_1,..,\sigma_s\}$ *is of the form* $\langle\texttt{mass value} = \kappa_k\rangle\psi_k\langle/\texttt{mass}\rangle$ *and* $\kappa_k = \frac{\Sigma\kappa_i^1 \times \kappa_j^2}{1 - \Sigma\kappa_n^1 \times \kappa_m^2}$ *such that* $\psi_k = \psi_i^1 \cap \psi_j^2$ *for the* (ψ_i^1, κ_i^1) *and* (ψ_j^2, κ_j^2) *pairs, and* $\psi_n^1 \cap \psi_m^2 = \emptyset$ *for the* (ψ_n^1, κ_n^1) *and* (ψ_m^2, κ_m^2) *pairs, and* ψ_k *is of the form* $\langle\texttt{massitem}\rangle\phi_{k_1}\langle/\texttt{massitem}\rangle, \ldots,$ $\langle\texttt{massitem}\rangle\phi_{k_z}\langle/\texttt{massitem}\rangle$.

The value $\kappa_\perp = \Sigma\kappa_n^1 \times \kappa_m^2$ (that is, $\Sigma_{A \cap B = \emptyset} (m_1(A) \times m_2(B))$) indicates how much of the total belief has been committed to the empty set while combining two pieces of uncertain information. A higher $\kappa_\perp$ value reflects either an inconsistency among the two sources or lower confidence in any of the possible outcomes from both sources.

Following the above example, if J Sky's PSA gives $PSA > 10$, then based on the XML document in Section 3, a new XML is generated for J Sky using his PSA value after integrating the method's reliability. This segment of the XML is

```
⟨belfunction⟩
  ⟨mass value = "0.455"⟩
     ⟨massitem⟩Cancer⟨/massitem⟩
  ⟨/mass⟩
  ⟨mass value = "0.245"⟩
     ⟨massitem⟩NoCancer⟨/massitem⟩
  ⟨/mass⟩
  ⟨mass value = "0.3"⟩
     ⟨massitem⟩Cancer⟨/massitem⟩
     ⟨massitem⟩NoCancer⟨/massitem⟩
  ⟨/mass⟩
⟨/belfunction⟩
```

Merging this BelVC with the one at the end of Section 4 using the procedure in Definition 7, we obtain a final diagnostic result with $m(\texttt{Cancer}) = 0.5609, m(\texttt{NoCancer}) = 0.3952$ and $m(\texttt{Cancer}, \texttt{NoCancer}) = 0.0438$ which strongly suggest that J Sky may have cancer.

6. Conclusion

A logical fusion framework that enables an easy modelling and merging of multiple summaritive and evaluative knowledge with uncertainty in XML, especially

in medical or bioscience domains, is reported in this paper. As XML is being increasingly used on the Web as a standard for data storage and exchange, modelling uncertain and incomplete information as well as merging these pieces of information have become an important and urgent issue. We believe our framework provides a formal platform for addressing these issues and has the potential to standardize the various proposals of modelling uncertain information in XML available so far.

References

1. P Ceravolo, E Damiani and B Oliboni. Fuzzy technique for medadata construction. *Proc. of IPMU'04*:1019-1026. 2004.
2. C Combi, B Oliboni, and R Rossato. Merging multimedia presentations and semi-structured temporal data: a graph-based model and its application to clinical information. *Artificial Intelligence in Medicine*, 2005.
3. J Cowie and W Lehnert. Information extraction. *Comm. of the ACM*, 39:81–91, 1996.
4. D Dubois and H Prade. *Possibility theory: An approach to the computerized processing of uncertainty.* Plenum Press, 1988.
5. N Elhadad, M Kan, J Klavans, and K McKeown. Customization in a unified framework for summarizing medical literature. *Artificial Intelligence in Medicine* 33,179-198, 2005.
6. M van Keulen, A de Keijzer and W Alink. A probabilistic XML approach to data integration. *Proceedings of ICDE'05*, 459-470, 2005.
7. A Hunter. Logical fusion rules for merging structured news reports. *Data and Knowledge Engineering*, 42:23–56, 2002.
8. A Hunter and W Liu. Fusion rules for merging uncertain information. *Information Fusion* (in press), 2005.
9. 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.
10. A Nierman and H Jagadish. ProTDB: Probabilistic data in XML. In *Proc. of VLDB'02*, LNCS2590: 646–657. Springer, 2002.
11. H Nottelmann and N Fuhr. pDAML+OIL: A probabilistic extension to DAML+OIL based on probability datalog. *Proc. of IPMU'04*:227-234, 2004.
12. T Pankowski and E Hunt. Data merging in life science data integration systems. *Intelligent Information Systems, Advances in Soft Computing*, Springer, 2005.
13. S Philippe and J Köhler. Using XML technology for ontology-based semantic integration of life science databases. *IEEE Trans. on Information Technology in Bioinformatics* 8(2):154-160, 2004.
14. G Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, 1976.