Inducing and Storing Generalised Evidences using Semantic Web formalisms

Razan Paul, Tudor Groza and Jane Hunter
School of ITEE, The University of Queensland, Australia
{razan.paul, tudor.groza, j.hunter}@uq.edu.au

Abstract

Over the course of the last decade, decision support systems have been used to assist clinicians and researchers in expanding the body of knowledge of particular (bio)-medical areas, as well as in diverse decision-making processes (e.g., diagnosis, treatment). Creating a decision support model (e.g., a rule base) requires a set of well-established medical guidelines built on mature domain knowledge. The absence of such mature domain knowledge has hindered the development of appropriate decision support methods in the skeletal dysplasia domain. In this paper, we make the first step towards providing a solution to this issue by proposing an ontology and associated extraction algorithm that can infer generalized evidences from existing bone dysplasia patient cases. This establishes the foundation for a decision support model based on evidential reasoning, which enables semi-automated diagnosis or key disease feature extraction.

Keywords: Evidence Ontology, Decision support methods, Evidential Reasoning

1 Introduction

Skeletal dysplasias are a heterogeneous group of genetic disorders affecting skeletal development. Currently, there are over 450 recognized bone dysplasias, structured in 40 groups. Patients with skeletal dysplasias have complex medical issues including short stature, bowed legs, a larger than average head and neurological complications. However, since most skeletal dysplasias are very rare (<1:10,000 births), data on clinical presentation, natural history and best management practices is sparse. Another reason for data sparseness is the small number of phenotypic characteristics typically exhibited by patients from the large range of possible phenotypic and radiographic characteristics usually associated with these diseases. Due to the rarity of these conditions and the lack of mature domain knowledge, correct diagnosis is often very difficult. In addition, only a few centres worldwide have expertise in the diagnosis and management of these disorders.

As there are no defined guidelines, the diagnosis of new cases relies strictly on parallels to past case studies.

Molecular genetics research on skeletal dysplasias has advanced considerably over the years – enabling the genetic defects responsible for more than 200 skeletal dysplasias to be identified. However, a lack of decision support methods and interoperable knowledge bases available to the skeletal dysplasia community – hinders collaborative diagnosis and research in this area. A rich knowledge base, together with associated decision support methods would enable researchers to verify known trends and to discover new, previously unknown trends among clinical attributes associated with this class of diseases, that can be used to assist and inform the decision making process associated with disease diagnosis and characterisation.

The general sparseness and disperse nature of skeletal dysplasia data has limited the development and availability of authoritative databases by the leading clinical and research centres. To make diagnoses, improve understanding and identify best treatments, clinicians need to analyse historical dysplasia patient data, verify known facts and relationships and discover new and previously unknown facts and relationships among the phenotypic, radiographic and genetic attributes associated with existing and new cases. For example, it is currently extremely difficult to recognise skeletal dysplasias that are etiologically related or to identify clinical or radiological characteristics that are indicative of defects resulting from a specific molecular pathway.

In order to do this, researchers and clinicians currently need to query many heterogeneous data sources and to effectively aggregate diverse types of data relating to phenotypic, radiographic and genetic observations. Given the appropriate data integration and reasoning tools, clinicians should be able to deduce, for example, that "mutations of the COL3A1 gene cause Platyspondylic lethal skeletal dysplasia which is characterised by short fingers in 90% of patients". However, this data integration step represents a significant challenge due to the extreme heterogeneity of the data models, metadata schemas and vocabularies, data formats and inconsistencies in naming and identification conventions.

Semantic Web standards (Berners-Lee, Hendler et al. 2001; Shadbolt, Hall et al. 2006; Allemang and Hendler 2008) encode and formalize data and background knowledge associated with a specific domain by means of standardized metadata schemas, controlled vocabularies and ontologies. These standards are critical to facilitating information sharing and integration. Hence, a key aim is to apply Semantic Web technologies to the data integration challenge described above by formalizing and modelling dysplasia data using ontologies and controlled vocabularies.

The above-mentioned issues also limit the potential of successfully applying existing or traditional knowledge representation, reasoning and decision support methods in the bone dysplasia domain, such as Rule Based Systems (Hudson 2006), Neural networks(Chan, Ling et al. 2011), Fuzzy cognitive maps (FCMs) (Hudson 2006; Gadaras and Mikhailov 2009; Papageorgiou, Papandrianos et al. 2009; Begum, Ahmed et al. 2010; Chan, Ling et al. 2011)) or Fuzzy Rule based classifications (Gadaras and Mikhailov 2009). Creating a decision support model (e.g., a rule base) requires a set of well-established medical guidelines and mature domain knowledge. However in the skeletal dysplasia domain, clinicians frequently have to diagnose patients with little or no similarity to past cases – this requires the generation of new evidence by combining existing evidence.

Our hypothesis is that by representing the knowledge and data using Semantic Web formalisms, and applying inductive reasoning on the resulting knowledge base – we can induce generalized evidences and store them in an Evidence Ontology. The result is an interoperable generalized evidence store for the skeletal dysplasia domain. Storing generalized evidences in an ontology enables sharing among and aggregation from multiple autonomous systems, thus leading to a distributed decision support approach.

The remainder of the paper is structured as follows: Section 2 provides a comprehensive overview of generically related efforts in the decision support area, while Section 3 introduces the fuzzy terminology used by our approach described in Section 4. Before concluding in Section 6, we briefly describe our evaluation plans in Section 5.

2 Related Work

Most prior work in representing generalized knowledge for medical decision support methods (Hudson 2006; Gadaras and Mikhailov 2009; Papageorgiou, Papandrianos et al. 2009; Begum, Ahmed et al. 2010; Chan, Ling et al. 2011) use some non-standard formalisms or proprietary formats which hinder integration, interoperability and efficient knowledge reasoning. They also lead to unjustified results by fusing all generalized knowledge into a black box system or assume the existence of mature domain knowledge. Moreover, some of these previous methods cannot evolve over time, due to their shallow knowledge representation formalisms. Case-based reasoning (Begum, Ahmed et al. 2010), on the other hand, cannot combine past evidences to form a new evidence for a given problem where no past similar evidence exists. This scenario is typical for rare diseases like skeletal dysplasias. It also uses non-generalized evidences, which does not guarantee correctness.

Rule based systems (Hudson 2006) and fuzzy rule-based classification (Gadaras and Mikhailov 2009) use exact matching on rules built on mature and established domain knowledge – which is inapplicable in a domain that suffers from data sparseness. The neural network approach (Chan, Ling et al. 2011) cannot provide justification for the resulting knowledge because it fuses all the evidence into the internal weights, whereas in the skeletal dysplasia domain, justification is very important to both clinicians and researchers in order to understand the underlying causal elements.

Today's decision support systems require the automatic integration of knowledge from multiple sources. However, the lack of interoperability and standard formalisms impede these systems to take advantage of the connectivity provided by the Web. Decision support systems (Goossen, IJntema et al. 2011; Lee and Wang 2011) using Semantic Web standards are being developed to overcome the above challenges. Semantic Web rule-based reasoning has been used for domain specific decision support methods, for example, in the Ambient Intelligence domain (Patkos, Chrysakis et al. 2010). However, such approaches cannot make use of underlying trends in instance data that have not been encoded as ontological background knowledge and cannot handle probabilistic uncertainties within the knowledge. Moreover, they cannot form new evidence by combining existing evidence via reasoning, where there exist no prior examples.

A recent related effort (Lee and Wang 2011) presents a novel fuzzy expert system for a diabetes decision support application using a 5-layer fuzzy ontology and a semantic decision support agent. However, as with its predecessors, this system also depends on mature and established domain knowledge, and uses fuzzy rule-based reasoning (Straccia 2008), which follows an exact matching approach.

Medical decision support systems have emerged from the co-evolution of research in decision support systems and medical informatics. In (Hussain, Abidi et al. 2007), a Semantic Web based Clinical Decision Support System is presented to provide evidence guided recommendations for follow-up after treatment for Breast Cancer. ControlSem (Andreasik, Ciebiera et al. 2010), a medical decision support system using Semantic Web technologies, was developed with the goal of controlling medical procedures. Similarly, in (Prcela, Gamberger et al. 2008), the authors present a medical expert system for heart failure. These expert systems use general purpose rule base reasoning (deductive reasoning) (Straccia 2008) because the underlying domain has well-defined rules and mature background knowledge.

Reasoning plays a vital role on the Semantic Web and is based on the background knowledge provided by the data model, logic and rules layers. Deductive reasoning is able to derive new knowledge, however, is relies completely on rules and existing ontological background knowledge and, hence, cannot make use of regularities in the instance data that have not been. In contrast, induction can exploit regularities in the instance data to discover new generalised rules or evidences.

Data mining is also applied to discover new information, hidden in patterns emerging from existing information. One of widely used techniques in data mining is finding association rules. The first pioneering work to mine conventional positive association rules using a level wise search algorithm was explained in (Agrawal, Imieli ski et al. 1993). Following this work, many other, improved, algorithms have been proposed, in particular for finding rules that represent decision occurring frequently based on a set of facts (Doddi,

Marathe et al. 2001; Pan, Li et al. 2005; Sheela and Shanthi 2009; Weng and Chen 2010).

Finally, Semantic Web Mining (Lisi 2006; Lisi 2006; Stumme, Hotho et al. 2006) is a new research area that aims to discover hidden knowledge from Semantic Web instance data by combining Semantic Web techniques and data mining. The newly discovered knowledge can then be used for enriching the domain model and, hence, possibly improve the future decision making process. Most of the existing work in Semantic Web mining applies existing data mining algorithms in the Semantic Web context. For example, (Lisi 2006) describes a middleware, SWing, to enable inductive reasoning on the Semantic Web. Similarly, (Maedche and Staab 2000) use association mining to extract relations from text.

3 Uncertainty, Fuzzy Set Theory and Ontology

Uncertainty comes in various forms: probabilistic uncertainty (e.g., "*There is a 65% chance I will get my promotion"*), vagueness (fuzziness – e.g., "*Mike is old to some degree*"), ambiguity, subjectivity, incompleteness, etc. It is widely accepted that uncertainty is an indispensable part of medical data, and that the first two types of uncertainty play an important role, e. g., a symptom may or may not occur with a disease, it has an uncertain relation with the disease, etc (De, Biswas et al. 2001; Straszecka 2006).

In the Semantic Web world, OWL ontologies and SWRL rules can be used to capture the domain knowledge in a highly expressive manner. However, these cannot model vague and uncertain knowledge, and implicitly concepts, such as "*short*" Limb, "*happy*" person or "*narrow*" chest, because are unable to capture the degree of happiness or the measure of shortness.

Fuzzy set theory and fuzzy logic (Ross 2010) are suitable formalisms to handle imprecise and vague knowledge of a particular domain. In traditional set theory, any element belongs or not to a set, in type-1 fuzzy set theory, any element can belong partially to a set. For example, Tim has "*short limb ≥ 0.5"* states that Tim has a short limb with a degree of at least 0.5. The traditional set theoretic operations are extended to the Fuzzy set and Fuzzy complement, union, intersection and the logical operation of implication are performed by special mathematical functions over the unit interval, and they are defined as fuzzy complement (c), tconorm (u), t-norm (t or $*$) and fuzzy implication ($\Rightarrow$) (Ross 2010) There are other uncertainties that type-1fuzzy set cannot handle, e.g., the *confidence* or *certainty* that Tim has "*short limb ≥ 0.5*". Type-2 fuzzy sets (Castillo, Melin et al. 2007) can handle these types of uncertainties by associating uncertainty with the membership function of a type-1 fuzzy set.

A fuzzy linguistic variable defines the terminology required to use a fuzzy concept like age in expressing rules and facts. A Fuzzy value is an instance of a fuzzy concept for a fuzzy linguistic variable, e.g., age is young >= 0.8. A Fuzzy linguistic Term is a word or expression used to facilitate the expression of Fuzzy value for a fuzzy linguistic variable. For example, age may have the Fuzzy terms {young, adult, old}. In our method, we employ type-2 fuzzy sets, fuzzy Linguistic variables, Fuzzy Value and Fuzzy Linguistic Term. Let's suppose we have to represent short limb with a degree of at least 0.8 and a certainty of 0.5. In this case we define a Fuzzy linguistic variable "limb", featured by three Fuzzy linguistic terms: {short, medium, long}. Our representation string will then be: 0.5 / (short, 0.8).

There are many concepts in medical domain that are vague and have no clear boundaries, such as "*young*", "*tall*" or "*small*". It is widely known that the crisp formalisms such as the one provided by OWL cannot handle vague and uncertain information on the Semantic Web. There are nevertheless, other ways to deal with such data. Firstly, it is possible to extend current Semantic Web languages to cope with fuzzy and uncertain information. Secondly, one can develop a specific, fuzzy, ontology. The World Wide Web Consortium (W3C) has set up a working group to work on representing and reasoning under uncertainty using ontologies. Results of this group can be seen in the existing fuzzy DL reasoners, like fuzzyDL (Simou and Kollias 2007; Bobillo and Straccia 2008) and FiRE (Simou and Kollias 2007). Straccia (Straccia 2010) and Pan (Pan, Stamou et al. 2007) have also described mechanisms the for persistent storage and querying of fuzzy and uncertain information in databases.

From the fuzzy ontology perspective, there have been several solutions proposed to date. Gu et al (Gu, Wang et al. 2007) describe a Fuzzy Ontology of edutainment based on reification of relations in OWL, a technique similar to representing n-array relations (Noy, Rector et al. 2006) in OWL. (Bobillo and Straccia 2009) propose an OWL ontology to represent important features of fuzzy OWL 2 statements, via temporary concepts (nodes) like ConjunctionConcept and ConceptAssertion. At a later stage, the same authors also propose a concrete methodology to represent a fuzzy ontology using OWL 2 annotation properties (Bobillo and Straccia 2010). Finally, (Stoilos, Stamou et al. 2005) extend OWL with fuzzy set theory in order to capture, represent and reason with fuzzy and uncertain information, while (da Costa, Laskey et al. 2008) propose the PR-OWL formalism by extending OWL to provide the ability to express probabilistic knowledge.

4 Research Methodology

Figure 1 presents the high level building blocks of our research methodology, represented by the SKELETOME ontology set, the Evidence Extraction Process and the Evidence ontology. In the following sections we detail each of the three building blocks.

Fig. 1. Inducing generalised evidences from the Skeletome Ontology set and storing them into the Evidence Ontology

SKELETOME Ontology set. The main role of the SKELETOME Ontology set is to improve the highly static and rigid format of the ISDS Nosology (Warman,

Cormier Daire et al.) by enabling a more flexible classification of the disorders and the integration with existing Web resources, such as the Human Phenotype Ontology (Robinson, Köhler et al. 2008) or the NCI Thesaurus (Sioutos, Coronado et al. 2007). The set is composed of three ontologies: the Bone Dysplasia ontology that captures the complex relations between the phenotypic, radiographic and genetic elements characterizing all skeletal dysplasias; the Patient ontology that models patient information and the Context ontology maintaining provenance information.

Evidence Ontology. The Evidence Ontology models uncertainty (both fuzzy / vagueness and probabilistic uncertainty) by re-using concepts from probabilistic uncertainty and from Fuzzy Theory, such as fuzzy value, fuzzy variable, fuzzy set, membership value and fuzzy term. It enables the representation of uncertain generalized evidences and helps to simplify uncertain knowledge representation in OWL. The crisp syntax of OWL DL is used within the Evidence ontology to enable the encoding of Fuzzy and probabilistic uncertainty semantics.

Evidence extraction process. The generalized evidence extraction from past patient cases stored via the SKELETOME ontology set is a crucial prerequisite for the implementation of any decision support method, like automated diagnosis or key disease feature inference. Without the extracted evidence, uncertainty reasoning cannot be performed. The actual extraction process uses Machine Learning techniques, and more specifically, a level wise search algorithm (Paul and Hoque 2010), to be able to infer evidences from the instances of the SKELETOME ontology concepts, made available by domain experts.

4.1 The SKELETOME Ontology Set

As already described, the SKELETOME Ontology Set consists of three ontologies that model together the skeletal dysplasia domain knowledge, patient information and context information.

The actual requirements of the ontology set emerged from the needs of the skeletal dysplasia community, and include the following:

Common terminology: The diagnosis and management of skeletal dysplasias depends on highly specialised domain knowledge across a number of disciplines (radiography, genetics, orthopaedics, physiotherapy), which is not easily comprehensible to individual communities or hospitals. In order to enable the exchange of knowledge between experts (across languages and disciplines), patients, their families and medical staff, a common terminology is required, hence leading to a shared conceptualisation of the domain.

Data integration: Large datasets containing rich information on molecules (genes, proteins) already exist and the information relevant to skeletal dysplasias needs to be extracted and cross-referenced with the clinical data and knowledge produced by SKELETOME. This requires integration both at conceptual level, as well as, at actual data / instance level.

Capturing provenance and expertise: The contributed content may take several forms, ranging from personal observations to scientific publications. Independently of the form, SKELETOME requires a mechanism to keep track of the provenance of the data and knowledge (to ensure proper privacy and access control), to provide a measure of certainty of derived data and to leverage expertise from the content and to streamline the delivery of the most relevant information / queries to the most appropriate person.

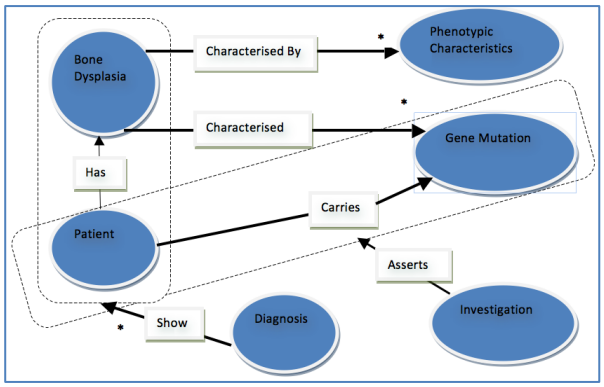

Fig. 2. Core concepts of the SKELETOME Ontology Set

Figure 2 depicts the core concepts of the SKELETOME Ontology Set. *Bone Dysplasia*, *Phenotypic Characteristic* and *Gene Mutation* are concepts defined by the Bone Dysplasia Ontology. The Bone Dysplasia ontology aims to complement the spectrum of existing ontologies and address the specific knowledge representation shortcomings of the ISDS Nosology (Warman, Cormier Daire et al.). None of the existing phenotype ontologies (e.g., the Human Phenotype Ontology) or well-known terminologies (e.g., SNOMED-CT) describe in detail skeletal dysplasias. As a result, our ontology provides a comprehensive, accurate and formal representation of the genotypes and phenotypes involved in skeletal dysplasias, together with their specific and disease-oriented constraints. As opposed to the current ISDS Nosology, this ontology enables a shared conceptual model, formalised in a machine-understandable language. In addition, it is continuously evolving and provides a foundational building block for facilitating further knowledge extraction and reasoning.

On the other hand, *Patient*, *Diagnosis* and *Investigation* are concepts present in the *Patient Ontology*. This ontology has the role to maintain patient data as instances of the domain knowledge, and in particular of associations of particular genotypic or phenotypic characteristics to different bone dysplasias. The graph created by the relationships between the above-mentioned concepts represents the input for the following steps of our research methodology.

4.2 Evidence Extraction from the SKELETOME Ontology set – the Level-wise algorithm

Figure 3 depicts the steps performed to create the decision support model. Firstly, data rows are extracted from the SKELETOME knowledge base. These are then transformed in the quantization process to make them suitable for comparison in the level-wise algorithm. Finally, the level-wise search algorithm is applied to

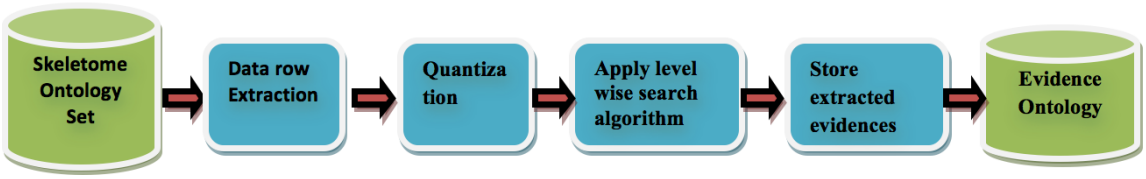

Fig. 3. Decision support model creation steps

discover generalised evidences, which are the stored as instances in the evidence ontology.

As a remark, the evidence extraction process assumes the strict use of positive statements in the data, due to the way in which patient cases (and clinical summaries) are described in this domain. More concretely, the SKELETOME ontology will only contain statements in the form of P implies Q, where P is a set of phenotypes and Q a skeletal dysplasia, without considering negation, e.g., P does not imply Q.

Another, different, remark needs to be made with respect to the evolution of the domain knowledge. The structure of the SKELETOME ontology will naturally evolve in accordance with the advances in the field. This evolution will be reflected both in the instance data (i.e., new patient cases), but also in the evidence extraction. From a technical perspective, we have currently plan to deal with this evolution by re-generating the evidences as part of a periodical batch process. However, for the future, we will consider incorporating such changes in the generalized evidences in an incremental manner.

Data Row Extraction. This step transforms the instance data present in the SKELETOME knowledge base, which is structured as interconnected graphs, into rows, as required by the level-wise search algorithm. Subsequently, it finds the most appropriate method to perform candidate and item set generation and to find rules within the given dataset, by also taking into account the physical resources associated with such a data-intensive method.

Algorithm: Discovering the trend of the instance data
Input: Data Rows from ontology, minimum support, minimum average certainity
Output : Itemsets which are strong candidates of Evidences.

1. K=1, S = {Ø};
2. Read the Knowledgebase about which attributes are Dysplasia (action type) and which are Symtoms(Non action).
3. I_k = Select all 1-itemsets which has support and average certainity greater or equal to minimum support and minimum certainity
4. While($I_k \neq Ø$) {
   4.1 K++;
   4.2 C_k = Candidate_generation(I_{k-1})
   4.3 CalculateCandidatesSupportAndCertainity(C_k)
   4.4 I_k = SelectDesiredItemSetFromCandidates (C_K, Sk , minimum support, minimum average certainity);
   4.5 S = S U S_k
5. return S

procedure Candidate_generation(I_{k-1})
1.For each Itemset $i_1 \in I_{k-1}$
  1.1 For each Itemset $i_2 \in I_{k-1}$
    1.1.1 Newcandidate, NC = Union(i_1,i_2);
    1.1.2 If Size of NC is k
      1.1.2.1 If NC contains one or no action item
        1.1.2.1.1 Add it to C_k if every subset of items is frequent.
2. return C_k;

Procedure CalculateCandidatesSupport(Ck)
1, For each transaction t of Data Rows
  1.1 CalculateSupportAndCertainityFromOneTransactionFor Cadidates(C_k, t);

procedure CalculateSupportFromOneTransaction ForCadidates(C_k, t)
1.C_t =Find the subsets of t which are candidate
2.For each candidate c € C_t
  2.1 c.count++
  2.2 Calculate Average Certainity

procedure SelectDesiredItemSetFromCandidates (C_K , S_k , minimum support, minimum average certainity)
1. For each Itemset c € C_k
  1.1 If c contains only non-action items
    1.1. 1 If c.support > = minimum support and c.certainity>= minimum average certainity
    1.1.2 Add it to I
  1.2 else if c contains action items
    1.2.1 If c.support > = minimum support and c.certainity>= minimum average certainity
    1.2.2 Add it to I & S_k
2. return I

Fig. 4. Trend discovery algorithm in instance data

Algorithm : Finding generalized evidences from desired item set.

Input: S(Desired item sets), minimum probabilistic uncertainty
Output: R (set of Evidences)
1. R = Ø
2. For each X Є S
   2.1 Symtom set P = (p_1, p_2,……. p_n){
   where p Є X and AC(p) ≠ 2}
   2.2 Dysplasia set Q = (q_1){
   where q Є X and AC(q) ≠ 1}
   2.3 CC = CaculateCorrelationCoefficient (P,Q)
   2.4 PU = CaculateProbabilisticUncertainty (CC,X.certainity)
   2.5 if PU <= minimum probabilistic uncertainty
   2.3.1 P↔Q is a valid evidence.
   2.3.2 R = R U (P↔Q)

Fig. 5. Generalized evidence discovery algorithm from the desired item set

Quantization. A second prerequisite to perform evidence extraction using the level-wise search algorithm is to transform the row data into a suitable format. Skeletal dysplasia data types can take multiple forms, ranging from categorical, or Boolean to continuous numerical data, interval, percentage or fraction. Continuous numerical data cannot be compared by direct difference as it may fail in recognizing some of the intrinsic data characteristics. For example, age intervals of equal width (e.g., $0 < age \leq 10$, $10 < age \leq 20$) may ignore certain data characteristics due to the ambiguous conventions associated with the patient's age interpretation, i.e., young, adult, or elder. A set of rules is created for each continuous numerical attribute using the knowledge of clinicians and researchers. A rule engine maps continuous numerical data to items using these developed rules. A domain dictionary is used to transform the data for discrete attributes.

Evidence extraction using Level-wise Search. A level wise search algorithm is developed to extract evidences from the SKELETOME knowledge base. The algorithm is based on the following statements:

- A statement (A->B) is treated as evidence based on the symmetric relationship strength between the antecedent and the consequent.
- Most generalised evidences involve a coherent subset of attributes, instead of implicitly including all possible attributes.
- Symptoms and observations lead to a particular decision and a decision can be a diagnosis or a procedure. All symptoms or observations are part of an antecedent and all diagnoses are part of a consequent.

Once the data has been transformed into a row-based format, the horizontal axis will represent patient instances. Fields composing the horizontal axis will be tagged as Action (representing the diagnosis) and Observation (representing symptoms, lab tests, genetic tests or radiographic features).

In the above-mentioned interval-based crisp quantization, elements near the boundaries of an interval will either be ignored or overemphasized(Kaya and Alhajj 2008). This may lead to loosing some of the underlying meaning of the data. For instance, an interval representing young persons might have a range between 18 and 40 years. In this instance, a person aged 17 would be a 0% representative and an 18 year old person would be 100%. However, the actual difference between these two ages is not that significant. This problem is caused by the sharp boundary between intervals(Kaya and Alhajj 2008). Implementing fuzziness can overcome this problem.

To address this issue, we use fuzzy quantization as an intermediate phase within the overall quantization step. For instance, we partition the values of the Age attribute into three fuzzy sets: low, medium and high. The intervals of low, medium and high could be {0−33}, {27−55} and {48−∞} respectively. In this instance, a person aged 30 years would be a representative of low with a certain degree and a representative of medium with a different degree. The domain experts define the corresponding fuzzy sets and their membership functions.

To have a clear understanding of the final data representation, below we present an example of a fuzzy encoding for a patient who exhibits three symptoms and has been diagnosed with a particular bone dysplasia:

Patient 1: {0.8/(s_1, 0.9), 0.8/(s_2, 0.9), 0.8/(s_3, 0.9), 0.8/(D, 0.9)}

s_1 = "Symptom X is Low"

s_2 = "Symptom Y is High"

s_3 = "Symptom Z is Low"

D = "Dysplasia BD is Medium"

Generalized evidence represents information inferred from generalized facts. The process of extracting generalized evidences from past patient cases consists of two steps:

1. Discovering the trend of the instance data by finding a desire item set using the level wise search algorithm.
2. Finding generalized evidences from the desired item set.

Step1: This step considers only the fuzzy terms of the fuzzy values, leaving out the membership value of these fuzzy terms. Even so, we consider only the fuzzy values that have membership value greater than a given

threshold (minimum membership value – *mmv*). At the same time, a fuzzy value with more than one fuzzy term will be converted into multiple transactions. For example: {Symptom X{Low, High}, Dysplasia X{High}} will be converted into: {Symptom X{Low}, Dysplasia X{High}} and {Symptom X{ High }, Dysplasia X{High}}.

Figure 4 details the trend discovery algorithm, where *K* is the size of the item set, *S* is set of the desired item set and $\{S_1, S_2, \ldots, S_y\}$ are the desired item sets of length {1,2, …, Y}. Also, $\{C_1, C_2, \ldots, C_q\}$ are the candidate item sets of length {1, 2, ..., Q} and $\{I_1, I_2, \ldots, I_t\}$ are frequent item sets of length {1, 2, …, t}.

Calculating Support and Average Certainty of an item set. If an item set has the items I = $\{i_1, i_2, i_3, \ldots, i_n\}$, there are *m* transactions in the knowledge base, we calculate the support and average certainty of the item set using the formulae below:

$$\text{Support or Probability(I)} = \frac{\sum_{k=1}^{m} I \in t_k}{total\ number\ of\ transactions}$$

$$\text{Average certainty (I)} = \frac{\sum_{t=1}^{m} \prod_{j=1}^{n} \text{certainty}(i_j)}{total\ number\ of\ transactions\ contains\ I}$$

Step 2: Figure 5 lists the algorithm for finding generalized evidences from the desired item set. Firstly, we reduce the desired item sets $\{S_2, S_3, \ldots, S_y\}$ only to those that have a skeletal dysplasia associated.

We then partition the symptoms and dysplasia of each item set into two sets: an action item set containing the dysplasias and a non-action item set containing the symptoms, with the symptom set of each of the initial item sets related to the dysplasia associated with the respective item set. Each of these relationships will represent generalized evidence. *AC* is the function that determines the type of an item, i.e., action or non-action.

AC(x) = 2 if it is non-action item/symtom

AC(x) = 1 if it is action item/dysplasia

Subsequently, we calculate the correlation coefficient between the action item set and the non-action item set of the evidence, and the probabilistic uncertainty by multiplying the resulting correlation coefficient and the average certainty. The generalized evidences having the probabilistic uncertainty value greater than a certain threshold will be considered as final result.

Ranking the generalized evidences could have been performed also by using *confidence*, which is another widely adopted interestingness metric. However, confidence does not account for the consequent frequency with the antecedent. In order to rank generalised medical evidences, we need a metric that takes into account frequency in both directions, i.e., the consequent frequency with the antecedent and the antecedent frequency with the consequent.

Correlation coefficient calculation. In a given medical relationship $s \rightarrow t$, s is a group of medical items where each item is constrained to appear in antecedent and t is a group of medical attributes where each item appears in consequent. Moreover $s \cap t = \emptyset$. For this relationship, the support is defined as support = $P(s,t)$ and the confidence is defined as = $P(s,t)/P(t)$, where P is the probability.

The correlation coefficient (also known as the Φ-coefficient) measures the degree of relationship between two random variables by looking at the degree of linear interdependency. It is defined by the covariance between the two variables divided by their standard deviations:

$$\rho_{st} = \frac{\text{Cov}(s,t)}{\sigma_s\, \sigma_t}$$

Here $\text{Cov}(s,t)$ represents the covariance of the two variables and σ_X and σ_Y stand for standard deviation. The covariance measures how two variables change together:

$$\text{Cov}(s,t) = P(s,t) - P(s)P(t)$$

Standard deviation is the square root of its variance and variance is a special case of covariance when the two variables are identical.

$$\sigma_s = \sqrt{\text{Var}(s)} = \sqrt{\text{Cov}(s,s)} = \sqrt{P(s,s) - P(s)P(s)} = \sqrt{P(s) - P(s)^2}$$

$$\sigma_t = \sqrt{P(t) - P(t)^2}$$

$$\rho_{st} = \frac{P(s,t) - P(s)P(t)}{\sqrt{P(s) - P(s)^2}\sqrt{P(t) - P(t)^2}}$$

$P(s,t)$ represents the support of an item set that consists of both s and t. Let the support of the item set be S_{st}. $p(s)$ and $p(t)$ will represent the support of antecedent s (S_s) and antecedent t (S_t), respectively. The value of S_{st}, S_s and S_t are computed during the desired item set generation. Using these values, we can calculate the correlation of every medical relationship among diverse groups of medical items. The correlation value will indicate the medical researcher how strong a medical relationship is from the perspective of historical data.

$$\rho_{st} = \frac{S_{st} - S_s\, S_t}{\sqrt{S_s - {S_s}^2}\sqrt{S_t - {S_t}^2}}$$

Hence, creating an association rule from the values of S_{st}, S_s and S_t provides us with a single metric, correlation coefficient, to represent the degree of relatedness between the antecedent and the consequent. For each medical relationship or rule, this metric is used to indicate the degree of relatedness between different groups of items. ρ_{st} takes values between -1 and +1. If two variables are independent then ρ_{st} is 0. When ρ_{st} is +1 the variables are considered perfectly positively correlated. A positive correlation represents an evidence of a general tendency of relatedness between a group of attribute values s and a group of attribute values y of a particular patient. The more positive the value is, the stronger the relationship is. When ρ_{st} is -1 the variables are considered perfectly negatively correlated.

Storing the Evidences. The fuzzy evidences discovered in the previous stage are stored in the Evidence Ontology

together with their corresponding probabilistic uncertainty.

4.3 Evidence Ontology

The Evidence Ontology (see Figure 6) is an OWL ontology we have built to represent uncertain generalised evidences. The ontology captures both fuzziness / vagueness and probabilistic uncertainty by re-using concepts from Fuzzy Theory, such as fuzzy value, fuzzy variable, fuzzy set, membership value, fuzzy term and probabilistic uncertainty. It enables the expression of uncertain information / evidence via ontological concepts and helps in simplifying knowledge representation in OWL.

The structure of the ontology comprises two conceptual layers:

- A fuzzy theory layer describing the features of fuzzy theory, such as, fuzzy term and membership value characteristics, and
- A conditional statements/conditional expressions layer modelling past skeletal dysplasia evidences with probabilistic uncertainty.

The SKELETOME ontology set is hence extended, via the Evidence Ontology, to enable analytical uncertainty reasoning on skeletal dysplasia cases. The resulting representation allows a powerful set of uncertainty operations while not introducing any inconsistency in the host ontology.

The syntax of the Fuzzy theory layer is based on OWL 2.0, while the semantics is based on the theory of fuzzy sets. Evidence are of the form antecedent → consequent with an associated probabilistic uncertainty, where antecedent consists of a set of Fuzzy Values of skeletal dysplasia symptoms, while the consequent is a set of Fuzzy Values of bone dysplasias. Range, domain, cardinality and functionality axioms are employed in the Evidence Ontology to keep the integrity of the Fuzzy theory, conditional statements and probabilistic uncertainty semantics. The object properties are exposed to establish association/relations between concepts, and data type properties are exposed to describe the attributes of concepts.

The Evidence Ontology has 8 main classes representing different concepts of fuzzy theory and conditional statements, listed below:

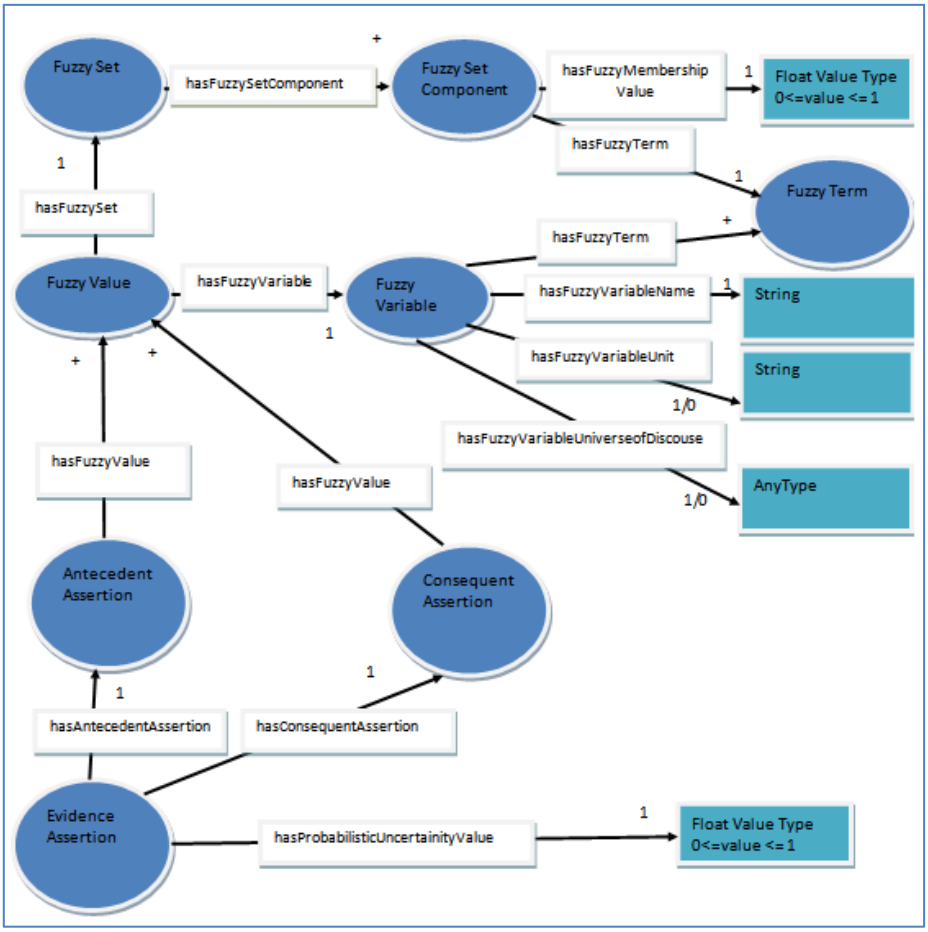

Fig. 6. Evidence Ontology concepts

Evidence Assertion represents evidences extracted from existing patient cases. An Evidence Assertion instance comprises an Antecedent Assertion, a Consequent Assertion and a Probabilistic Uncertainty.

Antecedent Assertion represents the antecedent part of an evidence. Its instances are composed of a set of fuzzy values.

Consequent Assertion represents the consequent part of an evidence. Its instances are composed of a set of fuzzy values.

Fuzzy Variable represents a fuzzy variable from the fuzzy theory. Fuzzy variables usually consist of a *name* (e.g., age), *terms* (e.g., child, young or blue), an *unit* (e.g., years) and the *universe of discourse* (e.g., 0-200).

Fuzzy Term represents a fuzzy term from the Fuzzy Theory, which is used as part of Fuzzy Set and Fuzzy Variable. For example, a fuzzy variable SymtomA may have fuzzy terms like Low, High or Medium. A Fuzzy Set may have the fuzzy term Medium with membership value 0.8. Each Fuzzy Term has a membership function.

Fuzzy Value represents the corresponding fuzzy value from the Fuzzy Theory and is value of a feature in a fuzzy sense instead of the crisp sense. After fuzzification, a numeric value converted in to Fuzzy Value and has two parts: a **Fuzzy Set** (i.e., the numeric value in the Fuzzy Terminology) and a **Fuzzy Variable** (Metadata about the Fuzzy terminology)

Fuzzy Set. Every member of a Fuzzy Set has membership degrees. A Fuzzy set instance is composed of a set of **FuzzySetComponents**.

FuzzySetComponent / FuzzySetMember represents a fuzzy element of the Fuzzy set theory. A FuzzySetComponent instance is composed of a fuzzy term and a corresponding membership value.

Currently, the Evidence Ontology has 8 classes, 10 object properties, 5 data type properties and no instances.

5 Evaluation Plans

To date, we have developed the Evidence ontology and identified a mechanism for inducing evidences. The next phase of the project involves evaluating these two aspects and refining/optimizing them based on the results.

Task-based evaluations (Porzel and Malaka 2004) will be used to assess the capability of the Evidence ontology to represent the generalized uncertain evidence. A set of use-cases, formulated as parameterized test questions and answer keys will be leveraged to characterize the ontology in terms of accuracy, insertion errors, deletion errors and substitution errors.

Similarly, to quantitatively assess the quality of the evidence extraction process, we will measure the evidence retrievability (recall) (Gupta, Fang et al. 2008) and the evidence spuriousness (precision) (Gupta, Fang et al. 2008). Evidence retrievability measures how well the underlying trends in past data have been discovered. Although retrievability provides a good estimate of the fraction of detected patterns in the data, it does not provide an estimate of the quality of the found patterns. The quality of a pattern is measured using spuriousness, which quantifies the number of items in the pattern that are not associated with the matching base pattern.

6 Conclusion

No prior research has investigated the induction of generalised evidences from immature domain knowledge, specifically in the skeletal dysplasia domain. This domain raises two important challenges with respect to developing decision support methods: (1) the absence of a wealth of background knowledge that would enable deductive reasoning, and (2) the sparseness of skeletal dysplasias data.

In this paper we have proposed a method for inducing generalized evidences from the existing patient cases, via a level-wise algorithm. The resulting knowledge is stored in the Evidence Ontology, which not only provides the foundational model for the development of appropriate decision support methods, but also a means for sharing generalised evidences in an interoperable manner.

Future work will focus on firstly evaluating the Evidence ontology and the level-wise algorithm. Following the evaluation and refinement of these components, the next step involves using instances of the Evidence ontology in conjunction with evidential reasoning, for automated diagnosis and key disease features extraction.

7 References

Agrawal, R., T. Imieli ski, et al. (1993). Mining association rules between sets of items in large databases, ACM.

Allemang, D. and J. A. Hendler (2008). Semantic web for the working ontologist: modeling in RDF, RDFS and OWL, Morgan Kaufmann.

Andreasik, J., A. Ciebiera, et al. (2010). ControlSem–distributed decision support system based on semantic web technologies for the analysis of the medical procedures. 3rd Conference on Human System Interactions (HSI), IEEE.

Begum, S., M. U. Ahmed, et al. (2010). "Case-Based Reasoning Systems in the Health Sciences: A Survey of Recent Trends and Developments." IEEE Transactions on Systems, Man, and Cybernetics--Part C: Applications and Reviews, IEEE.

Berners-Lee, T., J. Hendler, et al. (2001). "The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities." Scientific American 284(5): 34-+.

Bobillo, F. and U. Straccia (2008). fuzzyDL: An expressive fuzzy description logic reasoner, IEEE.

Bobillo, F. and U. Straccia (2009). "An OWL Ontology for Fuzzy OWL 2." Foundations of Intelligent Systems, Proceedings 5722: 151-160.

Bobillo, F. and U. Straccia (2010). "Fuzzy Ontology Representation using OWL 2." Arxiv preprint arXiv:1009.3391.

Castillo, O., P. Melin, et al. (2007). "Type-2 fuzzy logic: Theory and applications." Grc: 2007 Ieee International Conference on Granular Computing, Proceedings: 145-150.

Chan, K., S. Ling, et al. (2011). "Diagnosis of hypoglycemic episodes using a neural network based rule discovery system." Expert Systems with Applications.

da Costa, P. C. G., K. B. Laskey, et al. (2008). "PR-OWL: A Bayesian Ontology Language for the

Semantic Web." Uncertainty Reasoning for the Semantic Web I 5327: 88-107.

De, S. K., R. Biswas, et al. (2001). "An application of intuitionistic fuzzy sets in medical diagnosis." Fuzzy Sets and Systems 117(2): 209-213.

Doddi, S., A. Marathe, et al. (2001). "Discovery of association rules in medical data." Medical Informatics and the Internet in Medicine 26(1): 25-33.

Gadaras, I. and L. Mikhailov (2009). "An interpretable fuzzy rule-based classification methodology for medical diagnosis." Artificial Intelligence in Medicine 47(1): 25-41.

Goossen, F., W. IJntema, et al. (2011). News personalization using the CF-IDF semantic recommender, ACM.

Gu, H. M., X. Wang, et al. (2007). "Building a fuzzy ontology of edutainment using OWL." Computational Science - ICCS 2007, Pt 3, Proceedings 4489: 591-594.

Gupta, R., G. Fang, et al. (2008). Quantitative evaluation of approximate frequent pattern mining algorithms, ACM.

Hudson, D. L. (2006). Medical Expert Systems. Encyclopedia of Biomedical Engineering, John Wiley and Sons.

Hussain, S., S. R. Abidi, et al. (2007). "Semantic web framework for knowledge-centric clinical decision support systems." Artificial Intelligence in Medicine, Proceedings 4594: 451-455.

Kaya, M. and R. Alhajj (2008). "Online mining of fuzzy multidimensional weighted association rules." Applied Intelligence 29(1): 13-34.

Lee, C. S. and M. H. Wang (2011). "A Fuzzy Expert System for Diabetes Decision Support Application." Ieee Transactions on Systems Man and Cybernetics Part B-Cybernetics 41(1): 139-153.

Lisi, F. A. (2006). "A methodology for building Semantic Web Mining systems." Foundations of Intelligent Systems, Proceedings 4203: 306-311.

Lisi, F. A. (2006). "Practice of inductive reasoning on the Semantic Web: A system for Semantic Web mining." Principles and Practice of Semantic Web Reasoning 4187: 242-256.

Maedche, A. and S. Staab (2000). "Discovering conceptual relations from text." Ecai 2000: 14th European Conference on Artificial Intelligence, Proceedings 54: 321-325.

Noy, N., A. Rector, et al. (2006). "Defining n-ary relations on the semantic web." W3C Working Group Note 12.

Pan, H. W., J. Z. Li, et al. (2005). "Mining interesting association rules in medical images." Advanced Data Mining and Applications, Proceedings 3584: 598-609.

Pan, J. Z., G. Stamou, et al. (2007). Expressive querying over fuzzy DL-Lite ontologies, Citeseer.

Papageorgiou, E. I., N. Papandrianos, et al. (2009). "Fuzzy Cognitive Map Based Approach for Assessing Pulmonary Infections." Foundations of Intelligent Systems, Proceedings 5722: 109-118.

Patkos, T., I. Chrysakis, et al. (2010). "A Reasoning Framework for Ambient Intelligence." Artificial Intelligence: Theories, Models and Applications: 213-222.

Paul, R. and A. S. M. Hoque (2010). Mining irregular association rules based on action & non-action type data. Fifth International Conference on Digital Information Management (ICDIM) Thunder Bay, ON Canada, IEEE.

Porzel, R. and R. Malaka (2004). A task-based approach for ontology evaluation, Citeseer.

Prcela, M., D. Gamberger, et al. (2008). "Semantic web ontology utilization for heart failure expert system design." Studies in health technology and informatics 136: 851.

Robinson, P. N., S. Köhler, et al. (2008). "The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease." The American Journal of Human Genetics 83(5): 610-615.

Ross, T. J. (2010). "Fuzzy logic with engineering applications." Willy-India.

Shadbolt, N., W. Hall, et al. (2006). "The Semantic Web revisited." Ieee Intelligent Systems 21(3): 96-101.

Sheela, L. J. and V. Shanthi (2009). "DIMAR - Discovering Interesting Medical Association Rules form MRI Scans." Ecti-Con: 2009 6th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, Vols 1 and 2: 614-618.

Simou, N. and S. Kollias (2007). Fire: A fuzzy reasoning engine for impecise knowledge, Citeseer.

Sioutos, N., S. Coronado, et al. (2007). "NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information." Journal of Biomedical Informatics 40(1): 30-43.

Stoilos, G., G. Stamou, et al. (2005). Fuzzy OWL: Uncertainty and the semantic web, Citeseer.

Straccia, U. (2008). "Managing uncertainty and vagueness in description logics, logic programs and description logic programs." Reasoning Web: 54-103.

Straccia, U. (2010). An Ontology Mediated Multimedia Information Retrieval System, IEEE.

Straszecka, E. (2006). "Combining uncertainty and imprecision in models of medical diagnosis." Information Sciences 176(20): 3026-3059.

Stumme, G., A. Hotho, et al. (2006). "Semantic Web Mining - State of the art and future directions." Journal of Web Semantics 4(2): 124-143.

Warman, M. L., V. Cormier Daire, et al. "Nosology and classification of genetic skeletal disorders: 2010 revision." American Journal of Medical Genetics Part A.

Weng, C. H. and Y. L. Chen (2010). "Mining fuzzy association rules from uncertain data." Knowledge and Information Systems 23(2): 129-152.