DC Proposal: Decision Support Methods in Community-Driven Knowledge Curation Platforms

Razan Paul

School of ITEE, The University of Queensland,
St. Lucia, QLD 4072, Australia
razan.paul@uq.edu.au

Abstract. Skeletal dysplasias comprise a group of genetic diseases characterized by highly complex, heterogeneous and sparse data. Performing efficient and automated knowledge discovery in this domain poses serious challenges, one of the main issues being the lack of a proper formalization. Semantic Web technologies can, however, provide the appropriate means for encoding the knowledge and hence enabling complex forms of reasoning. We aim to develop decision support methods in the skeletal dysplasia domain by applying uncertainty reasoning over Semantic Web data. More specifically, we devise techniques for semi-automated diagnosis and key disease feature inferencing from an existing pool of patient cases – that are shared and discussed in the SKELETOME community-driven knowledge curation platform. The outcome of our research will enable clinicians and researchers to acquire a critical mass of structured knowledge that will sustain a better understanding of these genetic diseases and foster advances in the field.

Keywords: Decision Support Methods, Semantic Web, Skeletal Dysplasias.

1 Background and Problem Statement

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 centers 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.

Medical decision support systems can assist clinicians and researchers both in the research of skeletal dysplasias, as well as in the decision making process. However, the absence of mature domain knowledge and a lack of well documented, well

L. Aroyo et al. (Eds.): ISWC 2011, Part II, LNCS 7032, pp. 333–340, 2011.

structured past cases has hindered the development of decision support methods. Additionally, the general sparseness and disperse nature of skeletal dysplasia data has limited the development and availability of authoritative databases via the leading clinical and research centres. To make diagnoses, improve understanding and identify best treatments, clinicians need to analyze 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. In order to do this, they currently need to query many heterogeneous data sources and to effectively aggregate diverse types of data relating to phenotypic, radiographic and genetic observations. This 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.

The above-mentioned issues also limit the potential of successfully applying existing or traditional knowledge representation and decision support methods, such as Rule Based Systems [1], Neural networks [2], Fuzzy cognitive maps (FCMs) [3], Fuzzy Rule based classification [4] or Clustering algorithms [4], to the bone dysplasia domain. Creating a decision support model (e.g., a rule base) requires a set of well-established data models, data acquisition guidelines and mature domain knowledge. In this domain, clinicians have to diagnose patients with little or no similarity to past cases – this requires the generation of new evidence by combining existing evidence. This scenario combined with the general data scarcity issue has determined that distributed ontology reasoning [5] is a necessity within the skeletal dysplasia domain. Finally, because the study and understanding of skeletal dysplasias is still relatively immature, the justifications underpinning the decisions and the decision support methods, also need to be documented. All these elements make the knowledge representation and decision support methods in the bone dysplasia domain a very exciting and potentially productive area of research.

2 Aim and Objectives

Our hypothesis is that representing both the knowledge and the data via Semantic Web formalisms, together with the application of inductive and statistical reasoning on the resulting knowledge base, can support the development of efficient decision support methods in the skeletal dysplasia domain. More specifically, such approaches will enable the inferencing of key disease features from an existing pool of patients and the semi-automated diagnosis of the specific disease affecting new patients. This hypothesis can be further decomposed into the following research questions:

1. How can the generalized evidential statements (evidences), including their probabilistic uncertainties, be induced from existing patient cases stored in a Semantic Web knowledge base?
2. How can we efficiently build a comprehensive ontology that is capable of capturing the induced generalized evidences?
3. How can the probabilistic uncertainty of an evidential statement be incorporated to improve the precision of both the evidence learning (inductive reasoning) and the statistical reasoning?

4. How can the diagnosis of an undiagnosed skeletal dysplasia patient and the key features of a particular dysplasia be determined using the induced generalized evidential statements (including the associated justification)?
5. What is the optimum approach for combining existing skeletal dysplasia evidential statements to form new evidence?

In order to answer these research questions, this thesis will aim to achieve the following objectives:

a) The development of an ontology to store generalized evidential statements (evidences) that lay the foundation for further reasoning tasks.
b) Inducing the generalized knowledge (evidences) from the existing patient cases and encoding this knowledge in an interoperable manner.
c) Reasoning with this induced generalized knowledge to develop decision support methods. Instead of relying on a mature domain knowledge, or approach is to determine solutions directly from similar past examples. Our reasoning approach will utilize the generalized and inductive knowledge of past cases, concrete problem situations (new cases) and combine the learned knowledge from past cases to form new knowledge that will assist in solving specific problems.

3 Related Work

Existing online knowledge bases, such as the European Skeletal Dysplasia Network (ESDN) (www.esdn.org) and the Queensland Bone Dysplasia Registry (QBDR) (http://qbdc.org/bone-dysplasia-registry/) are ideal approaches for encouraging community-driven content exchange and curation. However, the underlying content is static, lacks formally defined semantics, and lacks decision support methods, thus making it difficult for the content to be reused, reasoned across and recombined for different purposes.

Most prior work in representing generalized knowledge for medical decision support methods [1-4, 6] 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 a mature established domain knowledge. Moreover, some of these previous methods cannot evolve over time, due to their shallow knowledge representation formalisms. Case-based reasoning [6], 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 [1] and fuzzy rule-based classification [4] 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 [2], 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.

It is widely accepted that uncertainty is an indispensable aspect of medical data. Bayesian reasoning [7], a widely used probability formalism, presents issues when

applied in this domain due to the estimation of the prior and conditional probabilities. Fuzzy sets are commonly used models to manage vagueness and imprecision in the medical domain [8]. Dempster-Shafer theory [9] is an alternative to representing probabilistic uncertainty mathematically. This is a potentially valuable tool to be used in the decision making process when precise knowledge is missing [9]. An important aspect of this theory is the combination of evidence obtained from multiple sources with the computation of a degree of belief that takes into account all the available evidence. Finally, as opposed to Bayesian reasoning, Dempster-Shafer theory does not require an estimation of the prior and conditional probabilities of the individual constituents of the set.

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 [10, 11] 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 [12]. 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 [11] 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 [13], 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 [14], a Semantic Web based Clinical Decision Support System is presented to provide evidence-guided recommendations for follow-up after treatment for Breast Cancer. ControlSem [15], a medical decision support system using Semantic Web technologies, was developed with the goal of controlling medical procedures. Similarly, in [16], the authors present a medical expert system for heart failure. These expert systems use general purpose rule base reasoning (deductive reasoning) [13] because the underlying domain has well-defined rules and a mature background knowledge.

4 Research Plan

We aim to develop decision support methods via an ontology-based interoperable framework tailored towards the skeletal dysplasia domain. This research combines ontological techniques with inductive and statistical reasoning techniques, and will be integrated within the SKELETOME[1] community-driven knowledge curation platform. Figure 1 presents the high level building blocks of the framework. The SKELETOME ontology[2], developed to capture the essential knowledge of skeletal dysplasia domain,

[1] http://itee.uq.edu.au/~eresearch/projects/skeletome/
[2] http://purl.org/skeletome/bonedysplasia

is a foundational prerequisite of our framework. While the SKELETOME ontology stores past and newly emerging patient cases, the Evidence ontology stores generalized evidences. Taking into account the lack of mature domain knowledge, the evidence extraction process induces generalized evidences from the existing patient cases and encodes the resulting knowledge in the Evidence ontology. Reasoning over the induced generalized evidences enables the development of the targeted decision support methods, i.e., automated diagnosis and identification of key disease features.

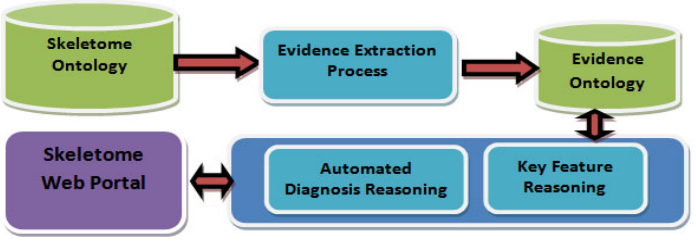

Fig. 1. Research methodology – building blocks

SKELETOME Ontology: The main role of the SKELETOME Ontology is to improve the highly static and rigid format of the ISDS Nosology [17], by enabling a more flexible classification of the disorders and integration with existing Web resources, such as the Gene Ontology, the Human Phenotype Ontology and the NCI Thesaurus. This ontology captures the complex relations between the phenotypic, radiographic and genetic elements that characterize all skeletal dysplasias.

Evidence Ontology: The Evidence Ontology models uncertainty (both fuzzy / 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 representation of uncertain generalized evidences and helps to simplify uncertain knowledge representation in OWL. OWL cannot encode Fuzzy and probabilistic uncertainty semantics. The crisp syntax of OWL DL will be used with the Evidence ontology to enable the encoding of Fuzzy and probabilistic uncertainty semantics.

Evidence Extraction Process: Generalized evidence extraction from past patient cases stored in the SKELETOME ontology is a crucial prerequisite for the implementation of the automated diagnosis and key feature inference capabilities. Without the extracted evidence, uncertainty reasoning cannot be performed. The actual extraction process will use Machine Learning techniques, and more specifically, a level wise search algorithm [18] that is able to infer evidences from the instances of the SKELETOME ontology concepts, made available by domain experts. The effectiveness of the method will be evaluated empirically. An *updating* module will be developed to ensure the continuous synchronization of the Evidence ontology instance base with the current patient repository.

Automated Diagnosis Reasoning: Clinicians can determine, to a level of approximation, possible diseases based on the medical symptoms that the patient presents. The vagueness (fuzziness) of medical symptoms is modelled by the

fuzzy set concept in the Evidence ontology. The probabilistic uncertainty of the medical fact is modelled by attaching a confidence/conditional probability value to each evidence. The automated diagnosis will infer a possible skeletal dysplasia based on a set of symptoms using Dempster–Shafer theory and fuzzy set theory (see Fig. 2). Generalized evidences are represented using the Evidence ontology, stored in the SKELETOME knowledge base and include probabilistic uncertainty values based on existing phenotype, radiographic and genetic information. The candidate hypothesis for an undiagnosed patient will be computed via reasoning. Dempster–Shafer theory will then be applied to the set of patient symptoms to determine the diagnosis based on the evidence stored in Evidence. In our case, the Dempster–Shafer calculation will only consider fuzzy terms, linguistic variables and probabilistic uncertainty, excluding the membership value of each fuzzy term.

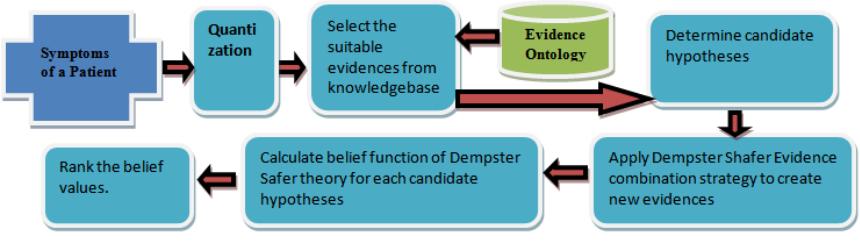

Fig. 2. Automated diagnosis reasoning based on a set of symptoms

Key Feature Reasoning: A key feature is a feature that is deemed highly characteristic of a specific skeletal dysplasia (for example, short fingers are a characteristic of Platyspondylic Lethal skeletal dysplasia). Inferring the key feature of a skeletal dysplasia from diverse phenotypic or genetic information is critical for the diagnosis of new patients. To determine such key features, our algorithm (depicted in Fig. 3) firstly determines the candidate hypotheses from the evidence stored in the Evidence Ontology. Then it applies domain-oriented ranking functions on the candidate hypotheses and presents the top K to the clinician using the system.

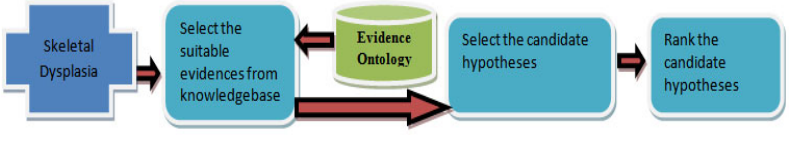

Fig. 3. Finding the Key Feature of a Dysplasia

5 Evaluation

The results of our research will be evaluated empirically. The patient cases used for evaluation will be collected from the ESDN and the QBDR - 90% of the cases will be used to construct our knowledge base, while the rest will act as test cases.

Evaluating the Decision Support Methods: The decision support methods will be evaluated on the basis of following criteria: (1) Decision-Making Effectiveness; (2)

Performance; (3) Ease of use and understating; (4) Scalability. Effectiveness will be measured using experimental methods combined with a usability study. Three metrics will be used to assess effectiveness: *Accuracy*, *Precision* and *Recall*. For the automated diagnosis aspect: *Accuracy* is the overall percentage of correctly diagnosed cases; *Precision* is the percentage of the correct diagnoses of a particular dysplasia; and *Recall* is the percentage of cases of a particular dysplasia that were correctly identified. For the key feature reasoning aspect: *Accuracy* is the overall percentage of correctly extracted features; *Precision* is the percentage of correctly selected key features and *Recall* is the percentage of dysplasias associated with a particular key feature that were correctly identified. To measure the performance of the decision methods, the following metrics will be used: Run time, Load time, Memory usage (main memory), and Memory usage (disk storage). The ease of use and understanding will be determined via a questionnaire with Likert-scale answers. We will define the testing environment and set of tasks to be performed by the participants. Observation data will also be collected from the usability study and used to complement to questionnaire. Scalability will measure how the methods scale as the number of patients in the KB increases. We will test our decision support methods against a number of different sized instance datasets and observe the changes in response times.

Evaluating the Evidence Ontology: Task-based evaluations [19] will be used to measure the generalized uncertain evidence representation capability of the Evidence ontology. 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.

Evaluation of the Evidence Extraction Process: To quantitatively assess the quality of the evidence extraction process, we will measure the evidence retrievability (recall) [20] and the evidence spuriousness (precision) [20]. 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 will be measured using spuriousness, which quantifies the number of items in the pattern that are not associated with the matching base pattern.

6 Conclusions

No prior research has investigated knowledge integration and decision support methods for the skeletal dysplasia domain - although it suffers from two important problems: (1) existing data is represented in a heterogeneous and non-interoperable manner, and (2) there are no mechanisms for building a consolidated and evolving knowledge base to support the decision making process. Our proposed research aims to address these problems by (1) using Semantic Web standards to formalize both the knowledge and data in the domain; (2) developing decision support methods based on past patient case studies, combined evidence and aggregated knowledge discovery. We believe that this research will advance the knowledge of the skeletal dysplasia community and expedite their understanding and diagnosis of skeletal dysplasias.

Acknowledgments. The work presented in this paper is supported by the Australian Research Council (ARC) under the Linkage grant SKELETOME - LP100100156.

References

1. Hudson, D.L.: Medical Expert Systems. In: Encyclopedia of Biomedical Engineering. John Wiley and Sons (2006)
2. Chan, K., et al.: Diagnosis of hypoglycemic episodes using a neural network based rule discovery system. Expert Systems with Applications (2011)
3. Papageorgiou, E.I., et al.: Fuzzy Cognitive Map Based Approach for Assessing Pulmonary Infections. In: Rauch, J., Raś, Z.W., Berka, P., Elomaa, T. (eds.) ISMIS 2009. LNCS, vol. 5722, pp. 109–118. Springer, Heidelberg (2009)
4. Gadaras, I., Mikhailov, L.: An interpretable fuzzy rule-based classification methodology for medical diagnosis. Artificial Intelligence in Medicine 47(1), 25–41 (2009)
5. Schlicht, A., Stuckenschmidt, H.: Towards distributed ontology reasoning for the web. IEEE (2008)
6. Begum, S., et al.: 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 (2010)
7. Wang, H.Q., Dash, D., Druzdzel, M.J.: A method for evaluating elicitation schemes for probabilistic models. IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics 32(1), 38–43 (2002)
8. Lekkas, S., Mikhailov, L.: Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases. Artificial Intelligence in Medicine 50(2), 117–126 (2010)
9. Dymova, L., Sevastjanov, P.: An interpretation of intuitionistic fuzzy sets in the framework of the dempster-shafer theory. Springer, Heidelberg (2010)
10. Goossen, F., et al.: News personalization using the CF-IDF semantic recommender. ACM (2011)
11. Lee, C.S., Wang, M.H.: A Fuzzy Expert System for Diabetes Decision Support Application. IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics 41(1), 139–153 (2011)
12. Patkos, T., Chrysakis, I., Bikakis, A., Plexousakis, D., Antoniou, G.: A Reasoning Framework for Ambient Intelligence. In: Konstantopoulos, S., Perantonis, S., Karkaletsis, V., Spyropoulos, C.D., Vouros, G. (eds.) SETN 2010. LNCS, vol. 6040, pp. 213–222. Springer, Heidelberg (2010)
13. Straccia, U.: Managing Uncertainty and Vagueness in Description Logics, Logic Programs and Description Logic Programs. In: Baroglio, C., Bonatti, P.A., Małuszyński, J., Marchiori, M., Polleres, A., Schaffert, S. (eds.) Reasoning Web. LNCS, vol. 5224, pp. 54–103. Springer, Heidelberg (2008)
14. Hussain, S., Raza Abidi, S., Raza Abidi, S.S.: Semantic Web Framework for Knowledge-Centric Clinical Decision Support Systems. In: Bellazzi, R., Abu-Hanna, A., Hunter, J. (eds.) AIME 2007. LNCS (LNAI), vol. 4594, pp. 451–455. Springer, Heidelberg (2007)
15. Andreasik, J., Ciebiera, A., Umpirowicz, S.: ControlSem–distributed decision support system based on semantic web technologies for the analysis of the medical procedures. In: 3rd Conference on Human System Interactions (HSI). IEEE (2010)
16. Prcela, M., Gamberger, D., Jovic, A.: Semantic web ontology utilization for heart failure expert system design. Studies in health technology and informatics 136, 851 (2008)
17. Warman, M.L., et al.: Nosology and classification of genetic skeletal disorders: 2010 revision. American Journal of Medical Genetics Part A (2010)
18. Paul, R., Hoque, A.S.M.: Mining irregular association rules based on action & non-action type data. In: Fifth International Conference on Digital Information Management (ICDIM). IEEE, Thunder Bay (2010)
19. Porzel, R., Malaka, R.: A task-based approach for ontology evaluation, Citeseer (2004)
20. Gupta, R., et al.: Quantitative evaluation of approximate frequent pattern mining algorithms. ACM (2008)