

Pronto: Probabilistic Semantic Web Ontologies

Pavel Klinov
(pklinov@clarkparsia.com)

Graduate Student / Intern
The University of Manchester, Clark & Parsia LLC

Agenda

- Semantic Web and Imprecision
- Probabilistic Ontologies
- One Real Life Application...

Semantic Web and Imprecision

Imprecision is Everywhere

- Uncertainty
 - 90% of birds fly. A bird is *likely* to fly
- Vagueness
 - “Retrieve all *inexpensive* hotels *nearby*”
- Subjectivity
 - “I *believe* that Tweety does not fly”
- Ambiguity
 - “Washington”. State? City? Team? President?

Imprecision in the Semantic Web

- Knowledge representation
- Data integration
 - Example: ontology alignment is usually uncertain
- Information and knowledge retrieval
 - Query processing
 - To what degree an object (Web page, text passage) matches my information need?
- Annotations

Uncertainty in Ontologies clark&parsia

- Taxonomic relationships can be uncertain
 - Birds *generally* have wings
 - “Automobile” *probably* means the same as “Car”
- Facts can be uncertain
 - Jim is Pete’s son with probability $>99\%$

Conceptual modeling and uncertainty

- Should be kept separated?
 - Strict domain knowledge provided via ontologies
 - Uncertainty handled elsewhere (rules, Bayesian networks, etc)
- Should be integrated?
 - Strict and uncertain domain knowledge mixed together in ontologies

Motivation

- We believe uncertainty should be allowed in ontologies. Why?
- Ontologies are *too useful* to abandon just because some statements are uncertain
 - Explicit modeling
 - Powerful and *explainable* reasoning
 - Lots of use cases

Real Life Application

Breast Cancer Risk Assessment

- Problem statement: estimate risk of developing breast cancer given set of risk factors:
 - Age
 - Ethnicity (e.g., being an Ashkenazi Jew is a risk)
 - Medical history (whether immediate relatives have BRC)
 - Genetics (e.g., BRCA1, BRCA2 gene mutations)

Breast Cancer Risk

- Risks

- Absolute (short term, lifetime)
- Relative (increased, decreased)

- Examples:

- An average US woman has 12% risk of developing BRC in her lifetime
- Family history of BRC doubles the risk

BRC and Uncertainty

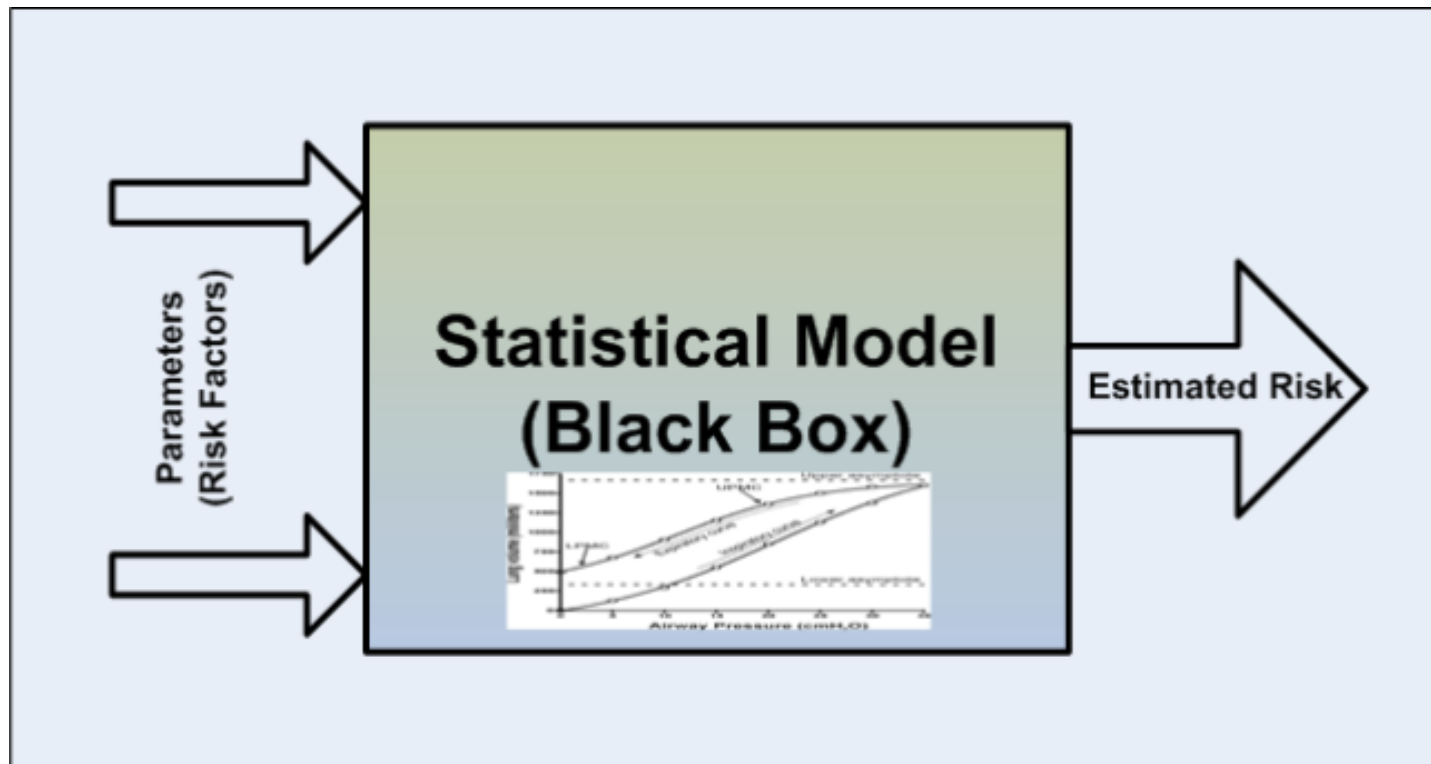
- The domain is highly uncertain:
 - Not all risk factors are known
 - Not all risk factors have been sufficiently investigated (e.g., gene mutations)
 - Not all relationships between risk factors are sufficiently studied
 - Presence of certain factors cannot always be precisely determined => errors (e.g., unreliable medical history)

BRC and Uncertainty

- These issues complicate use of **NCI Thesaurus** – one of the largest formal ontologies in the world!
- Solutions:
 - Abandon it and use statistical models (Gail model)
 - Or enrich it with uncertain statements?

BRC and Statistical Modeling

- One approach is to use a “black-box” statistical modeling



BRC and Statistical Modeling

- Advantages:
 - Fast calculation
 - Can handle large number of parameters
 - Can be statistically validated
- Disadvantages
 - Non-transparent
 - Adding new factors isn't always easy
 - Multiple models are hard to integrate

BRC and Ontological Modeling

- One alternative is to re-use existing medical ontologies (NCI Thesaurus)
- Key idea: reduce risk assessment to ontological reasoning
 - What is probability that Jane belongs to the concept “WomenBRCInNext10Years”
- Here we need a *probabilistic reasoner*.
That's Pronto!

Probabilistic Ontologies

Probabilistic Uncertainty clark&parsia

- Probability theory is the best studied theory of managing uncertainty
- Well suited for capturing:
 - Statistics
 - Degrees of belief
- Natural question: can semantics of uncertain statements in ontologies be captured *probabilistically*?

Probabilistic Ontologies

- Classical ontologies augmented with probability assertions:
 - Uncertain subclass relationships. Birds are flying objects with probability >0.9
 - Uncertain individual statement. Tweety is a flying object with probability <0.05

Probabilistic Ontologies clark&parsia

- Classical ontological languages (OWL) do not provide built-in constructs for the representation of such statements
- Language requirements:
 - Syntactic constructs for representing conditional statements
 - Well-defined formal semantics
 - Plausible entailment relations

What we are doing

- We picked a formalism
 - Expressive
 - Integrates with existing methodologies
 - Provides good reasoning services
- Implemented it
 - Reasoner (Pronto – extension to Pellet)
 - Browser (OWLSight)
- Evaluated it

Formalisms

- Two major options:
 - Bayesian Networks
 - Probabilistic Logic
- We picked the latter. Because:
 - Better integration with OWL (DL)
 - Better expressivity
 - Higher level of formality (e.g., soundness of reasoning)

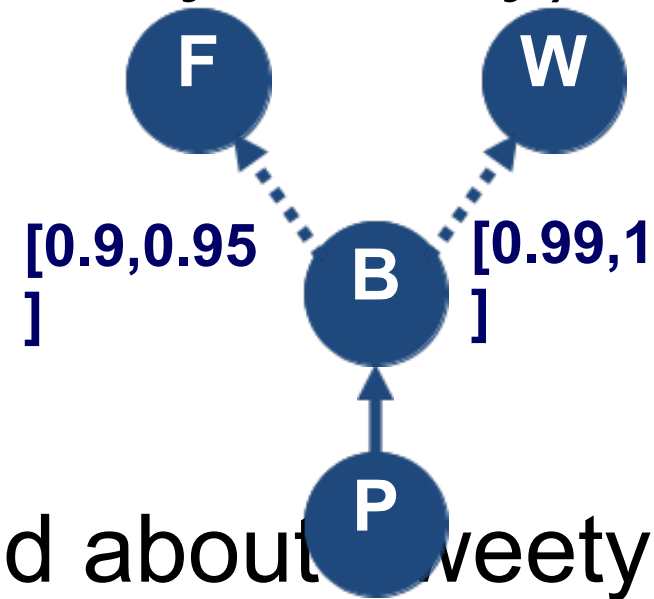
- Probabilistic Ontology:
 - Classical part (OWL)
 - Probabilistic part (conditional constraints)
- Conditional constraint:
 - Expression of the form $(D|C)[l,u]$
 - C, D – OWL classes, $[l,u]$ – interval
- Example:
 - $(\text{Fly}|\text{Bird})[0.9,0.95]$

Expressivity

- Default statistical knowledge
 - (Flies|Bird)[0.9,0.95] – *generally* birds fly with probability between 90% and 95%
- Degrees of belief
 - (Flies)[0,0.05] – Tweety flies with probability less than 5%

Reasoning

- Sound (agrees with Probability Theory)
- Allows exceptions:
 - (Flies|Bird)[0.9,0.95]
 - (Wings|Bird)[0.99, 1.0]
 - (Flies|Penguin)[0,0.5]
 - Tweety is a Penguin
- What should be concluded about Tweety's ability to fly? Having wings?



Back to Breast Cancer

Modeling

- Classical part – OWL ontology (NCI Thesaurus can be imported)
- Probabilistic part
 - Statistical knowledge about risk factors. (IncreasedRisk|MotherAffected)[0.4,0.5]
 - Individual risk factors. (HighLevelOfEstrogen) [0.7,0.8] for Jane

Modeling

- What can be captured:
 - How presence of risk factors affects the risk
 - Relations between risk factors
 - Examples: Ashkenazi Jews are more susceptible to gene mutations. Estrogen and progestin are known to strengthen negative impact of each other

Risk Assessment

- Pronto computes risk as probabilistic entailment
- Risk = probability that a woman is an instance of a particular OWL class
- Example:
 - Jane:(WomenBRCInNext10Years)[0,0.05]

Explanations

- Pronto can *explain* the results
 - Helen, you're in top risk category because your mother was affected before the age of 60 and you have high level of estrogen
- It filters out all irrelevant statistics and factors

Conclusions

Advantages

- Transparency
 - Ontology is the model (visualization, explanations, etc)
- Extensibility
 - Classical part can be extended
 - Probabilistic statements can be added
- Semantic rigidity (entailments based on logical proofs)

Issues and Challenges

- User perspective
 - Modeling methodologies need to be developed
 - Probability should be well understood by modelers
 - Syntactic sugar and interfaces need to be added

Issues and Challenges

- Technical perspective
 - Scalability!
 - Evaluation methodologies and test suites to be developed
 - Better explanation services
 - More reasoning services
 - Independence assertions

Work in progress

- Scalability
 - *Optimizing Pronto*
- Modeling methodologies
 - Gathering use cases, feedback
- Experimental methodologies
 - Developing test ontologies, benchmark test suites
- Attracting interest!

Questions? Comments?