UNIVERSITÀ DEGLI STUDI DI PISA

FACOLTÀ DI SCIENZE MATEMATICHE FISICHE E NATURALI
CORSO DI LAUREA SPECIALISTICA IN INFORMATICA

TESI DI LAUREA

# Modelli di integrazione di reti bayesiane ed ontologie: analisi e confronto

CANDIDATO
Massimiliano Brocchini

RELATORE
Professor Franco Turini

CONTRORELATORE
Professor Andrea Corradini

Anno Accademico 2009/2010

# Indice

# Elenco delle figure

# Capitolo 1

# Introduzione

Attualmente le ontologie rappresentano lo standard per modellare la conoscenza riguardante concetti e relazioni fra di essi all'interno di uno specifico dominio. Utilizzate fin dalla fine degli anni settanta per formalizzare l'insieme della conoscenza in grado di descrivere un dominio di interesse, le ontologie sono state definite dal Word Wide Web Consortium (W3C) lo standard per rappresentare la conoscenza semantica riguardante le entità presenti nel Semantic Web, al fine di consentire l'interoperabilità semantica fra fonti di dati, pagine web e applicazioni software create in maniera indipendente, e rappresentate in maniera disomogenea.

Tuttavia, le ontologie sono tipicamente basate su una logica esatta e nessuno dei linguaggi creati per definirle e per ragionare su di esse è in grado di gestire informazioni incerte o parziali riguardanti il dominio modellato. Informazioni incerte o parziali esistono durante ogni fase di sviluppo e utilizzo di un'ontologia: dalla creazione del modello, al popolamento delle istanze, fino al ragionamento. Adesso che le ontologie sono considerate una tecnologia matura e sono applicate in campi disparati — siti web sociali, medicina e biologia, "*digital libraries*", "*data mining*" — è chiaro che si cerchi di poter rappresentare informazioni in evoluzione o parzialmente definite, quindi incerte, all'interno delle ontologie.

La teoria delle probabilità è una scelta naturale per far fronte all'incertezza delle informazioni da trattare: integrando le ontologie con la teoria delle probabilità si rende possibile, non solo applicare le tecniche di ragionamento fornite dalle ontologie ad informazioni incerte o incomplete, ma anche eseguire nuove forme di ragionamento probabilistico sul modello e sui dati a disposizione.

Applicare forme di calcolo probabilistico al contenuto di un'ontologia senza riuscire a tradurre nella teoria probabilistica i legami fra i vari elementi che la compongono, non permetterebbe di

ottenere i risultati desiderati. Infatti, troppa informazione verrebbe persa ignorando la struttura dei dati modellata nell'ontologia. Citando Glenn Shafer: "*Probability is not really about numbers; it is about the structure of reasoning*". Un modello per la rappresentazione della conoscenza in termini probabilistici, in grado di rappresentare la struttura dei legami dei dati su cui ragionare è costituito dalle reti bayesiane, un formalismo grafico in grado di codificare in maniera compatta i legami di influenza fra i concetti modellati ed il grado di fiducia associato al verificarsi di un certo evento (la probabilità). Oltre a permettere di tradurre i legami presenti fra gli elementi dell'ontologia in legami probabilistici, le reti bayesiane hanno il vantaggio, rispetto ad altri modelli di rappresentazione dell'incertezza come la logica fuzzy, di essere già utilizzate, sia nella ricerca che nell'industria, per la modellazione ed il ragionamento in diversi domini in cui trovano applicazione anche le ontologie come ad esempio la biologia e la medicina.

## 1.1 Motivazione dell'analisi svolta

La scelta di analizzare e confrontare lo stato dell'arte sui modelli di integrazione fra reti bayesiane e ontologie nasce dalla constatazione dell'emergere di numerose pubblicazioni sull'argomento a partire dagli anni duemila. Nonostante il gran numero di pubblicazioni trovate non è stato possibile individuare una metodologia standard per l'integrazione di tutta la conoscenza modellata da un'ontologia con una rete bayesiana. Durante lo studio dei lavori analizzati è emerso chiaramente come l'unico fattore comune ai vari modelli sia quello di partire da una modellazione esatta della conoscenza, attraverso un'ontologia, per andare ad aggiungere, successivamente, la conoscenza probabilistica. Ogni modello analizzato sceglie un sottoinsieme degli elementi dell'ontologia da modellare nella rete bayesiana tralasciando informazioni sulla struttura o sugli individui che potrebbero essere indispensabili in alcuni domini applicativi. Inoltre, questa differenza fra gli elementi modellati nella rete bayesiana si ripercuote, anche, nelle forme di ragionamento probabilistico messe a disposizione: da un lato perché cambiano gli "oggetti" su cui eseguire il ragionamento, dall'altro perché alcuni modelli offrono varianti di ragionamento più strutturato e complesso di quello normalmente messo a disposizione da una rete bayesiana grazie alla modellazione scelta per la rete bayesiana.

In definitiva, quest'analisi nasce dall'impossibilità di dichiarare un modello di integrazione migliore, in assoluto, degli altri e dalla mancanza di pubblicazioni che confrontino in dettaglio diversi modelli di integrazione, mostrando pregi e limiti di ognuno e che analizzino le differenze indipendentemente dal dominio di applicazione dell'ontologia.

## 1.2 Struttura della tesi

Questa tesi è organizzata come segue. Il Capitolo 2 offre una presentazione sintetica dei concetti introduttivi: la Sezione 2.1 fornisce un'introduzione alle ontologie spiegando cosa siano le ontologie, di quali elementi si compongano, come possano essere specificate tramite due formalismi del W3C e quali forme di ragionamento offrano; la Sezione 2.2 introduce il concetto bayesiano di probabilità evidenziandone le differenze dalle definizioni frequentista e classica di probabilità, procede, poi, definendo le caratteristiche delle reti bayesiane e quali forme di ragionamento implementino. Il Capitolo 3 offre un'analisi dei modelli scelti per essere presentati in questa tesi, fornendo la spiegazione dei punti salienti dei metodi di integrazione, di alcune delle loro caratteristiche e spiegando alcuni termini propri di ogni metodo utili alla comprensione del confronto descritto nel Capitolo 4. Il Capitolo 4 mette a confronto le differenze dei vari modelli partendo dai requisiti che ciascuno ha per poter essere applicato, fino alle forme di ragionamento probabilistico messe a disposizione. Infine, il Capitolo 5 presenta le conclusioni di questo lavoro.

# Capitolo 2

# Concetti introduttivi

## 2.1 Ontologie

### 2.1.1 Definizione

Il termine "ontologia" nasce in filosofia, dove indica lo studio dell'essere. Esistono diverse definizioni, più o meno equivalenti, del termine ontologia nel contesto dell'informatica. Una buona definizione è quella data da Tom Gruber [8]: "*un'ontologia definisce un insieme di primitive di rappresentazione con cui modellare un dominio di conoscenza o discorso. Le primitive di rappresentazione sono tipicamente classi, attributi e relazioni. Le definizioni delle primitive di rappresentazione includono informazioni sul loro significato e vincoli affinché la loro applicazione sia logicamente consistente[...]*". Una ontologia è dunque una descrizione formale degli oggetti e relazioni fra di essi, che appartengono ad uno specifico dominio.

### 2.1.2 Elementi di un'ontologia

La conoscenza di un dominio può essere specificata in una ontologia utilizzando quattro componenti:

- Classi: un insieme astratto, o collezione di oggetti. Può contenere individui o altre classi
- Attributi: specificano proprietà, fatti generali riguardo agli elementi di una classe
- Relazioni: un'interazione fra classi o individui

- Individui: oggetti membri degli insiemi definiti dalle classi

Un'ontologia non deve necessariamente contenere degli individui, ma uno degli scopi generali di una ontologia è fornire un sistema per classificare individui, anche se questi non appartengono esplicitamente all'ontologia.

All'interno di una ontologia è possibile notare una chiara distinzione fra la conoscenza generica rispetto al dominio di interesse e la conoscenza specifica per un particolare problema. Questa distinzione causa la divisione dell'ontologia in due componenti: *TBox* e *ABox*.

La *TBox* ("*Terminological Box*") contiene le informazioni sul dominio di interesse, sotto forma di conoscenza relativa alle classi, alle relazioni fra di esse e agli attributi.

La *ABox* ("*Assertional Box*") contiene le informazioni su uno specifico problema (inteso come instanziazione particolare del discorso sul dominio), sotto forma di conoscenza sugli individui e sulle relazioni fra di essi.

### 2.1.3 Resource Description Framework (RDF)

RDF[1] è un linguaggio per la rappresentazione di informazioni riguardanti risorse disponibili sul Web definito dal Word Wide Web Consortium (W3C). Il meccanismo di rappresentazione offerto da RDF è quello delle triple "soggetto","predicato","oggetto" (il predicato viene chiamato anche proprietà della tripla). L'insieme delle triple forma un *grafo RDF*: l'insieme dei nodi del grafo è l'insieme dei soggetti e oggetti delle triple, mentre i predicati costituiscono gli archi. Questo meccanismo permette di esprimere semplici affermazioni sulle risorse del dominio utilizzando proprietà identificate da un nome e dotate di un valore. Perché queste affermazioni possano essere definite formalmente è necessario che venga definito il vocabolario che può essere utilizzato per specificare che le affermazioni riguardano specifiche classi di risorse e che verranno utilizzate specifiche proprietà nella descrizione delle risorse oggetto del dominio modellato. RDF Schema (RDFS)[2] permette di definire il vocabolario associato alle triple RDF restringendo l'insieme di termini validi per la definizione dei componenti delle triple. Tuttavia RDFS non permette di catturare molte relazioni fra classi e proprietà a causa di un vocabolario non sufficientemente ricco, inoltre ciascuna tripla forma un' asserzione distinta: l'aggiunta di una qualsiasi tripla non modifica il significato delle triple esistenti.

[1] http://www.w3c.org/RDF
[2] http://www.w3.org/TR/rdf-schema/

Per far fronte a queste difficoltà il W3C ha definito la famiglia di linguaggi *OWL* come estensione del vocabolario di RDF.

### 2.1.4 Web Ontology Language (OWL)

Il linguaggio OWL[12] (Web Ontology Language) è il linguaggio di markup correntemente definito come standard dal W3C per la definizione e instanziazione di ontologie. OWL è una estensione sintattica e semantica di RDF. OWL estende il potere espressivo di RDF e le possibilità di ragionamento e inferenza sulle relazioni fra concetti e proprietà, tramite tre sotto linguaggi di espressività e complessità crescente:

- OWL Lite offre una gerarchia di classificazione (tramite la relazione di sotto classe "*is_a*" anche espressa come "*rdfs:subClassOf*") e la definizione di semplici vincoli (es. la cardinalità può prendere come valori solamente 0 o 1)
- OWL DL è progettato per offrire il massimo dell'espressività mantenendo completezza e decidibilità computazionali dei sistemi di ragionamento applicabili. Rappresenta la variante OWL più diffusa e utilizzata
- OWL Full è il linguaggio più espressivo della famiglia ed è stato creato per mantenere una compatibilità diretta con RDFS. Non fornisce alcuna garanzia computazionale e permette ad un'ontologia di estendere il vocabolario predefinito in cui è stata codificata. La semantica di OWL Full non è compatibile con quella di OWL DL e OWL Lite

Esempio di sintassi OWL: definizione di classi, sotto classi, proprietà e restrizioni

```
<owl:ObjectProperty rdf:ID="hasParent">
  <rdfs:domain rdf:resource="#Animal"/>
  <rdfs:range rdf:resource="#Animal"/>
</owl:ObjectProperty>
<owl:Class rdf:ID="Person">
  <rdfs:subClassOf rdf:resource="#Animal"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#hasParent"/>
```

```
        <owl:toClass rdf:resource="#Person"/>
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction owl:cardinality="1">
        <owl:onProperty rdf:resource="#hasFather"/>
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#name"/>
        <owl:minCardinality>1</owl:minCardinality>
     </owl:Restriction>
    </rdfs:subClassOf>
</owl:Class>
```

### 2.1.5 Interrogazioni e forme di ragionamento

La formalizzazione della struttura e delle relazioni fra i concetti del dominio codificata in una ontologia, permette di utilizzare le ontologie per: estrarre i dati inseriti, cioè recupero delle istanze e delle loro informazioni in maniera analoga alle interrogazioni delle basi di dati relazionali, e per inferire nuova conoscenza tramite forme di ragionamento automatico che rendano esplicita parte dell'informazione contenuta implicitamente nell'ontologia.

È possibile utilizzare un'ontologia per rispondere ad interrogazioni simili a quelle utilizzate nelle basi di dati relazionali:

- "*instance checking*": stabilire se una particolare istanza appartiene ad un dato concetto
- "*relation checking*": stabilire se una relazione diretta esiste fra due istanze o classi
- recupero di informazioni: dato un concetto, recuperare tutte le istanze appartenenti a quel concetto

Le interrogazioni sopra elencate sfruttano solamente la conoscenza codificata esplicitamente nell'ontologia e non coinvolgono forme di ragionamento automatico. Interrogazioni più complesse richiedono l'applicazione di forme di ragionamento deduttivo, tipico delle "*description logic*" messe a disposizione dalle ontologie OWL DL.

- soddisfacibilità: stabilire se una descrizione di classe C è necessariamente vuota (insoddisfacibile) o no [3]
- sussunzione: stabilire se una sussunzione $C \sqsubseteq D$ è conseguenza logica (delle descrizioni delle classi) dell'ontologia
- equivalenza: stabilire se un'equivalenza fra classi è conseguenza logica dell'ontologia
- disgiunzione: stabilire se l'insieme degli individui appartenenti a due classi è soddisfacibile

Combinando le tipologie di ragionamento sopra elencate con un sistema di inferenza, che applichi "*forward chaining*" e/o "*backward chaining*" a seconda dei compiti richiesti al motore inferenziale, è possibile: estrarre nuova conoscenza da quella contenuta nell'ontologia (realizzazione del *modus ponens* tramite applicazione del "*forward chaining*"), verificare la validità e consistenza dell'ontologia stessa, rispondere alle interrogazioni di un utente (tipicamente tramite "*backward chaining*"). Molti sistemi software permettono inoltre di definire un insieme di regole da far applicare al motore inferenziale per consentire di modificare l'ontologia stessa, sia per quanto riguarda gli elementi della *TBox* che della *ABox*, (es. Sezione 3.2) o per definire vincoli, aggiuntivi a quelli definiti nell'ontologia, da applicare durante l'estrazione di conoscenza.

## 2.2 Reti bayesiane

### 2.2.1 Concetto bayesiano di probabilità

La probabilità bayesiana viene interpretata come il grado di fiducia di un individuo nel verificarsi di un dato evento [11]. Mentre la probabilità frequentista assegna probabilità ad eventi casuali so-

[3] non si tratta di verificare l'esistenza di un individuo nell'ABox appartenente a C, ma stabilire se esiste un modello logico in cui l'insieme degli elementi che soddisfano C non sia necessariamente vuoto.

lamente secondo la frequenza delle loro occorrenze[4], la probabilità bayesiana permette di assegnare un valore di probabilità a qualsiasi tipo di enunciato.

Secondo Pearl [11], contrariamente alla tradizione di definire le probabilità condizionate $P(A|B)$ in termini di probabilità congiunte $P(A,B)$, per i filosofi bayesiani la relazione condizionata è la vera base di partenza per il ragionamento da cui derivare la probabilità di eventi congiunti tramite la formula:

$$P(A,B) = P(A|B)P(B) \tag{2.1}$$

Sotto questo punto di vista, $B$ indirizza verso una parte della conoscenza e $A|B$ specifica un evento $A$ nel contesto specificato da $B$ (es. un sintomo $A$ nel contesto di una malattia $B$).
La probabilità che si verifichi un singolo evento $A$ è definita dalla somma pesata della probabilità di tutte le forme in cui $A$ possa manifestarsi:

$$P(A) = \sum_i P(A|B_i)P(B_i) + P(A|\neg B_i)P(\neg B_i) \tag{2.2}$$

Nell'ambito di questo lavoro di analisi dei metodi di integrazione fra reti bayesiane e ontologie è necessario ricordare, come sottolineato da Pearl, che ogni formula di calcolo delle probabilità (come l'Equazione (2.2)), deve essere sempre intesa applicata ad un contesto più ampio, che definisce le assunzioni considerate di senso comune o dominio del nostro discorso (es. l'equiprobabilità della frequenza di uscita di una faccia di un dado non truccato). Pertanto l'Equazione (2.2) non è altro che una notazione abbreviata per l'equazione

$$P(A|K) = \sum_i P(A|B_i,K)P(B_i|K) + P(A|\neg B_i,K)P(\neg B_i|K) \tag{2.3}$$

Un'importante generalizzazione dell'Equazione (2.1) è la formula chiamata "*chain rule*". Dato un insieme di $n$ eventi $E_1, E_2, \ldots, E_n$la probabilità dell'evento congiunto $(E_1, E_2, \ldots, E_n)$ può essere scritto come il prodotto di $n$ probabilità condizionate applicando ripetutamente l'Equazione (2.1) in un ordine qualsiasi:

$$P(E_1, E_2, \ldots, E_n) = P({}_nE|E_{n-1}, \ldots, E_2, E_1) \ldots P(E_2|E_1)P(E_1) \tag{2.4}$$

[4] Analogamente la definizione classica di probabilità permette di assegnare probabilità solamente agli eventi di cui è possibile stimare il numero di casi favorevoli e quelli possibili.

Nell'inferenza su reti bayesiane (Sezione 2.2.3) la "*chain rule*" giocherà un ruolo fondamentale nel calcolo delle distribuzioni di probabilità di un evento.

L'inferenza della probabilità, sempre intesa come grado di fiducia, di una ipotesi $H$ una volta che si è manifestata una prova $e$ può essere calcolata come aggiornamento del grado di fiducia che era stato accordato all'ipotesi $H$ prima del verificarsi di $e$ tramite la formula dell'inversione di Bayes:

$$P(H|e) = \frac{P(e|H)P(H)}{P(e)} \tag{2.5}$$

dove $P(H)$ rappresenta il grado di fiducia in $H$ prima di essere a conoscenza di $e$, cioè la *probabilità a priori*, mentre $P(e)$ è solamente un fattore di normalizzazione. $P(H|e)$ viene chiamata *probabilità a posteriori*.

### 2.2.2 Definizione e caratteristiche

Una rete bayesiana è una grafo orientato aciclico in cui:

- i nodi rappresentano variabili aleatorie
- gli archi rappresentano relazioni di dipendenza condizionata
- ad ogni nodo è associata una tabella di probabilità condizionata (*CPT*)

Il grafo orientato aciclico di una rete bayesiana è una rappresentazione compatta delle proprietà di dipendenza e indipendenza dell'intera distribuzione di probabilità congiunta modellata dalla rete.

La struttura del grafo permette l'esplicitazione delle relazioni di dipendenza e di definire localmente ad ogni nodo le distribuzioni di probabilità condizionata che collegano un nodo ai suoi genitori. Questo porta ad una naturale scomposizione della distribuzione di probabilità congiunta, nel prodotto di probabilità condizionate locali di ciascun nodo, infatti, la "*chain rule*" espressa nell'Equazione (2.4) permette di calcolare la distribuzione di probabilità congiunta nel seguente modo:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|parent(X_i)) \tag{2.6}$$

Se un nodo $X_i$ non ha genitori la sua distribuzione di probabilità si dice non condizionata o a priori.

Le relazioni di indipendenza condizionata all'interno del grafo della rete bayesiana sono definite tramite il criterio di d-separazione ("*d-separation*"):

*un cammino non orientato* $\pi = < u, \ldots, v >$ *in un grafo orientato aciclico* $G = (V, E)$*, è detto bloccato da* $S \subseteq V$ *se* $\pi$ *contiene un nodo w tale che si verifichi almeno una delle seguenti condizioni:*

- $w \in S$ *e gli archi di* $\pi$ *non si incontrano testa a testa in w*
- $w \notin S \wedge \{discendenti(w)\} \cap S = \emptyset$ *e gli archi di* $\pi$ *si incontrano testa a testa in w*

*Per tre sottoinsiemi A,B,S di V, A e B sono d-separati da S, se tutti i cammini non orientati fra A e B sono bloccati da S.*

Una volta identificate le variabili che danno evidenza dei fatti noti, sfruttando il criterio di d-separazione per calcolare dinamicamente la distribuzione di probabilità congiunta è possibile ridurre il numero dei termini di quest'ultima anche in modo esponenziale. Ad esempio Nir Friedman in un seminario tenuto presso la "Association for the Advancement of Artificial Intelligence (AAAI)" nel 1998[5] ha mostrato come una rete bayesiana composta da 37 variabili booleane avesse solamente 509 parametri coinvolti nel calcolo della distribuzione di probabilità congiunta invece dei $2^{37}$parametri che sarebbero risultati dall'applicazione dell'Equazione (2.4) senza tener conto della struttura della rete.

Utilizzando la rete bayesiana rappresentata nella Figura 2.1 come esempio, è possibile calcolare il numero di parametri necessario per il calcolo della distribuzione di probabilità congiunta della rete:

- Chain rule:
  $P(C,R,S,W) = P(C)P(R|C)P(S|R,C)P(W|R,C,S)$ : numero di termini = 2*4*8*16 = 1024
- Utilizzando la proprietà di indipendenza condizionata:
  $P(C,R,S,W) = P(C)P(R|C)P(S|C)P(W|R,S)$ : numero di termini = 2*4*4*8 = 256

Nella rete rappresentata l'attivazione dell'irrigatore è indipendente dalla pioggia e il fatto che l'erba sia bagnata è indipendente dal cielo nuvoloso.

[5] http://www.cs.huji.ac.il/~nir/AAAI98-Tutorial/Tutorial.htm

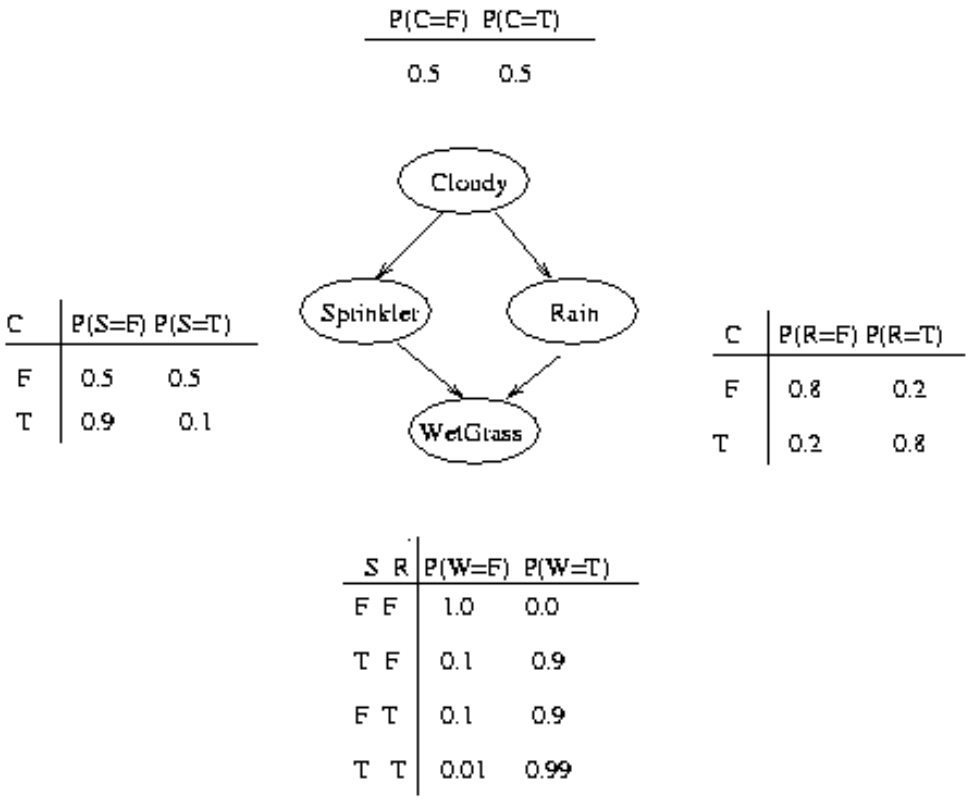


Figura 2.1: Esempio di rete bayesiana

### 2.2.3 Inferenza su reti bayesiane

Esistono tre schemi di inferenza sulle reti bayesiane: ragionamento causale, ragionamento diagnostico e "*explaining away*"

Lo schema del ragionamento causale permette di calcolare la probabilità di un effetto E data una delle sue cause C (Figura 2.2). Il ragionamento causale si basa sulla propagazione in avanti di informazioni dai padri verso i figli. Nell'esempio di Figura 2.2, è possibile calcolare:

$$P(C|A) = P(C|A,B)P(B) + P(C|A,\neg B)P(\neg B)$$

Lo schema del ragionamento diagnostico permette di calcolare la probabilità di una causa C dato il suo effetto E (Figura 2.3). Il ragionamento diagnostico si basa sulla propagazione all'indietro di

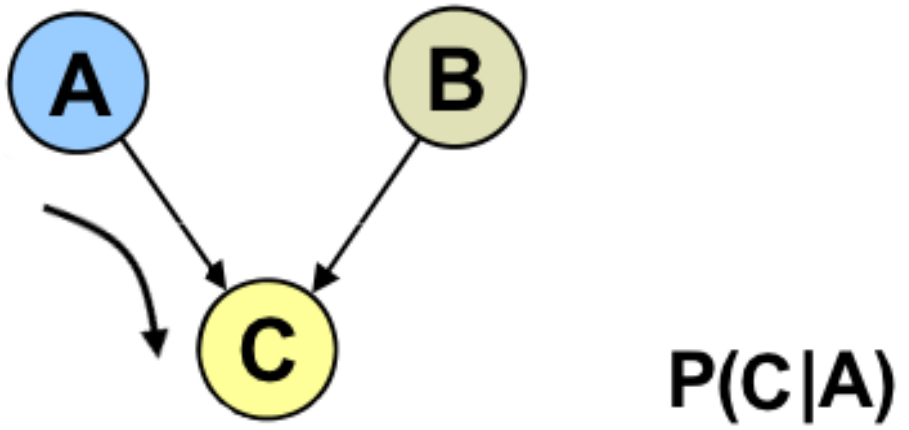


Figura 2.2: ragionamento causale

informazioni dai figli verso i padri. Applicando il teorema di Bayes, il ragionamento diagnostico viene trasformato in ragionamento causale a meno di un fattore di normalizzazione. Nell'esempio di Figura 2.3, è possibile calcolare:

$$P(A|C) = P(C|A)\frac{P(A)}{P(C)}$$

P(A|C) è stato definito con l'applicazione della formula di Bayes evidenziando la componente calcolabile applicando il ragionamento causale P(C|A), e il fattore di normalizzazione P(A)/P(C).

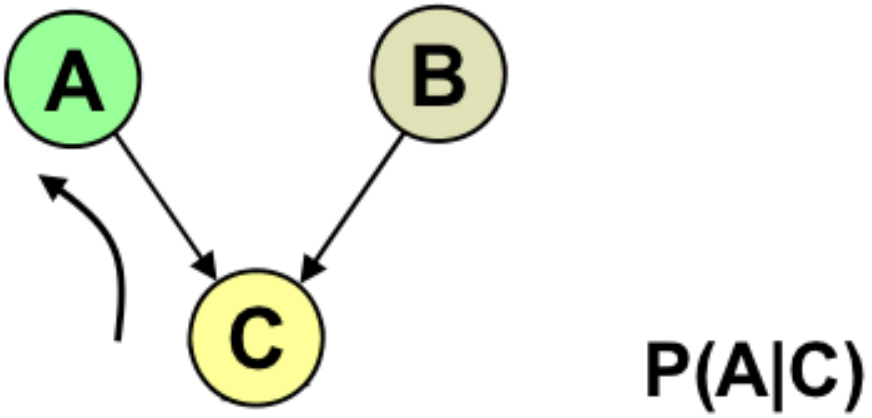


Figura 2.3: ragionamento diagnostico

Infine lo schema dell'*explaining away* permette di calcolare la probabilità di una causa C, relativa ad un effetto E, data un'altra causa C' di E (Figura 2.4). L'*explaining away* combina una passo di ragionamento causale dentro un processo di ragionamento diagnostico.

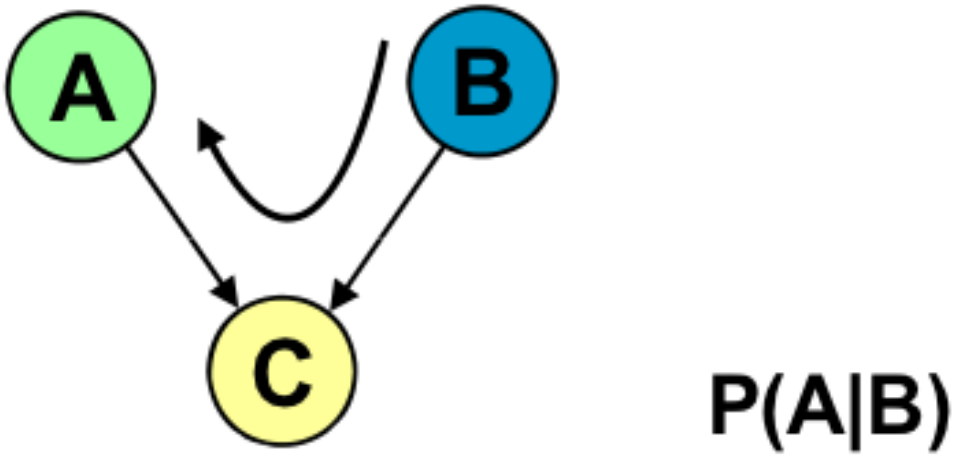


Figura 2.4: "*explaining away*"

In generale l'inferenza probabilistica, sia esatta che approssimata, su una rete bayesiana è NP-hard. Esiste una tipologia di reti in cui l'inferenza è polinomiale nel numero dei nodi, quella dei "*polytree"*: dato un grafo orientato aciclico, questo è un "*polytree*" se è privo, anche, di cicli non orientati.

# Capitolo 3

# Analisi della letteratura

In questo capitolo verranno analizzati i modelli di integrazione fra reti bayesiane ed ontologie scelti per essere presentati in questa tesi. L'analisi è volta a mostrare come gli autori abbiano definito i propri modelli di integrazione e questi siano notevolmente differenti gli uni dagli altri. Il capitolo ha, anche, lo scopo di introdurre alcuni concetti e vocaboli specifici per i singoli modelli che saranno ripresi nel Capitolo 4 e di presentare alcuni punti di forza e di debolezza, o mancanze, che emergono durante la trattazione delle differenti strategie di integrazione proposte.

## 3.1 OntoBayes

Calmet e Yang [14, 13] propongono un estensione di OWL, chiamata OntoBayes, che integra l'inferenza di una rete bayesiana all'interno di una ontologia con lo scopo di facilitare l'uso del ragionamento probabilistico nel processo di prendere decisioni (*decision making*).

L'integrazione di una rete bayesiana in una ontologia $O$ avviene tramite due passi di estensione della rappresentazione di $O$ in OWL:

1. annotazione delle probabilità
2. annotazione delle dipendenze

### 3.1.1 Annotazione delle probabilità

Le probabilità, che devono essere stimate da un esperto del dominio, vengono annotate direttamente nell'ontologia di partenza arricchita da tre nuove classi OWL:

1. PriorProb
2. CondProb
3. FullProbDist

Le prime due classi identificano le probabilità a priori e quelle condizionate. Entrambe le classi hanno una "*datatype property*" che esprime il valore numerico della probabilità.
La classe "*FullProbDist*" modella le distribuzioni di probabilità e ha due *"object property"* disgiunte "*hasPrior*", "*hasCond*" che la legano alle classi "*PriorProb*" e "*CondProb*" rispettivamente.

### 3.1.2 Annotazione delle dipendenze

In OntoBayes i nodi della rete bayesiana derivante dall'ontologia di partenza rappresentano solamente *object property* o *datatype property*[1].

Le relazioni di dipendenza fra i nodi della rete bayesiana devono essere costruite esplicitamente aggiungendo alla proprietà OWL dominio della relazione di dipendenza un elemento RDF "*<rdfs:dependsOn>*".
Es. la Figura 3.1 mostra come l'object property "*Customer - Buy - Product*" dipenda dalla datatype property "*Price*" della classe "*Product*".

### 3.1.3 Rete bayesiana risultante

Dalle relazioni di dipendenza descritte nella Sezione 3.1.2 è possibile costruire il grafo orientato aciclico rappresentante la struttura della rete bayesiana. Le annotazioni delle probabilità descritte nella Sezione 3.1.1 costituiscono le tabelle della probabilità condizionata e l'insieme delle probabilità a priori.

[1] Non vengono modellate relazioni fra classi

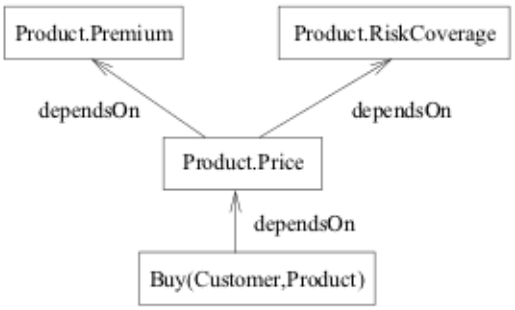


Figura 3.1: dipendenze fra proprietà (Fonte: [13], pag.8, fig.2)

La definizione esplicita delle relazioni di dipendenza fra variabili aleatorie permette di collegare anche quelle classi che nella ontologia apparterrebbero a due domini distinti e separati (es. "*Customer*" e "*Natural_disaster*" in Figura 3.2).

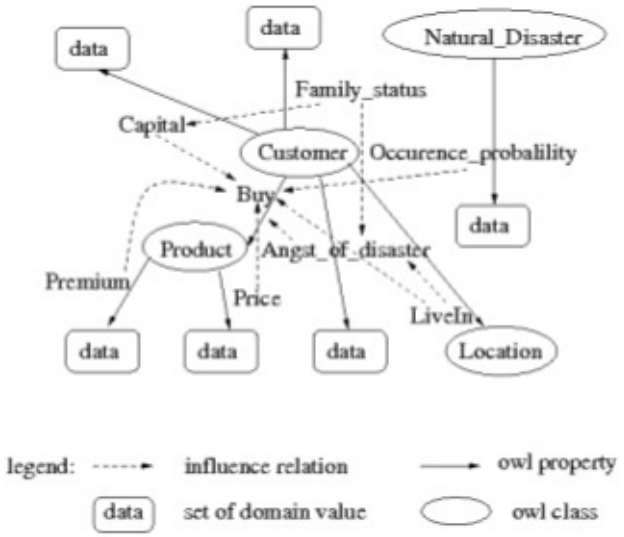


Figura 3.2: Modello grafico della versione estesa di OWL (Fonte: [14], pag.6, fig.10)

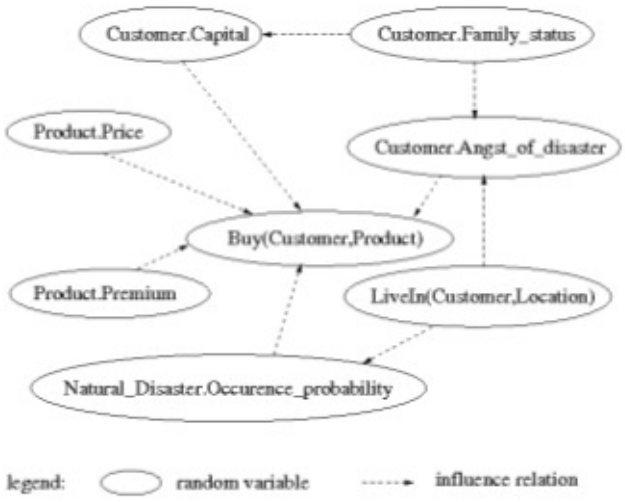


Figura 3.3: Rete bayesiana corrispondente al modello OWL di Figura 3.2 (Fonte: [14], pag.6, fig.11)

## 3.2 Constructing bayesian networks automatically using ontologies

Il modello di OntoBayes [14, 13] (Sezione 3.1) permette di aggiungere inferenza probabilistica ad una ontologia modellata con OWL. Tuttavia, il procedimento descritto ha i difetti di:

1. richiedere la presenza di un esperto di dominio che abbia conoscenze adeguate sulle reti bayesiane per specificare correttamente le relazioni di dipendenza fra proprietà e le relative distribuzioni di probabilità.
2. modificare l'ontologia di partenza

Devitt et al. dichiarano in [4] di definire un sistema che permetta di eliminare il difetto numero due e di ridimensionare, almeno in parte, il primo.

Il lavoro presentato è ambizioso perché propone un metodo generico per la definizione di uno schema per la trasformazione di una ontologia in una rete bayesiana. I passi descritti per la creazione di una rete bayesiana sono i seguenti:

1. identificazione delle variabili di interesse

2. specifica degli attributi dei nodi della rete bayesiana

3. creazioni degli archi fra variabili

4. assegnazione di una distribuzione di probabilità condizionata

### 3.2.1 Identificazione delle variabili di interesse e specifica degli attributi dei nodi della rete bayesiana

Detta "*domain ontology*" l'ontologia OWL-DL da tradurre in rete bayesiana, gli autori di [4] definiscono un'ontologia ausiliaria "*BN ontology*" con concetto radice *"BNnode*" per specificare i concetti appartenenti a "*domain ontology*" che devono essere tradotti in variabili aleatorie. Tutti e soli i concetti che ereditano da "*BNnode*" e i loro discendenti nella "*domain ontology*" vanno a formare una terza ontologia che verrà tradotta nella rete bayesiana.

L'ontologia risultante dalla fusione della "*domain ontology*" e della "*BN ontology*" può essere estesa con nodi concetto (un "*bm:BehaviourModelNode*" e alcuni "*bm:ConceptXNode*") che specificano ulteriori restrizioni al processo di creazione della rete bayesiana. Questa ontologia estesa viene chiamata "*Behaviour Model Ontology"* (Figura 3.4).
Il concetto "*BehaviourModelNode*" definisce gli attributi che tutti i nodi della rete bayesiana dovranno avere per il particolare dominio di applicazione da rappresentare.
I concetti "*bm:ConceptXNode*" definiscono i valori degli attributi dei nodi corrispondenti alle istanze della classe "*ConceptX*" della "*domain ontology*" che devono essere trattate in maniera particolare (es. specificare intervalli di discretizzazione per variabili continue, restrizioni sugli insiemi di valori di alcune "*datatype property*") (Figura 3.5).

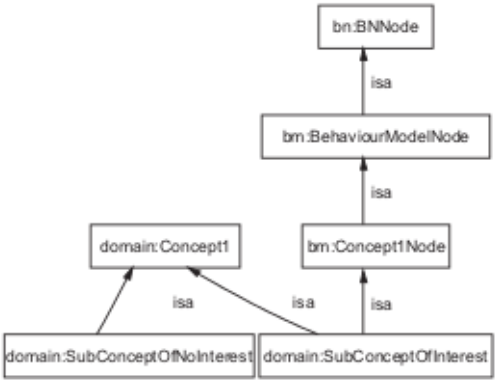


Figura 3.4: Struttura della "*Behaviour Model Ontology*" (Fonte: [4], pag.5, fig.3)

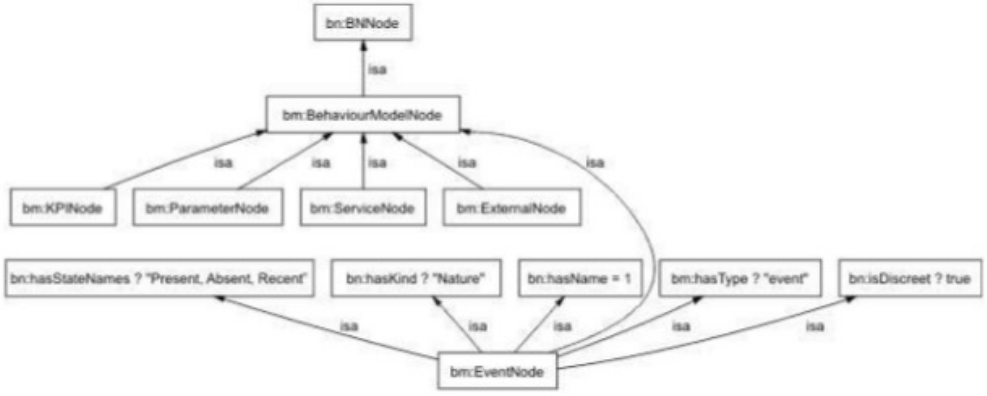


Figura 3.5: Restrizioni sulle "*datatype property*" per il nodo "*Event*" (Fonte: [4], pag.6, fig.4)

### 3.2.2 Creazione degli archi fra variabili

Per costruire gli archi della rete bayesiana un esperto del dominio deve definire nel sistema di inferenza dell'ontologia "*BN ontology*" un insieme di regole ad hoc per il dominio trattato. Questo permette di collegare fra loro variabili corrispondenti a concetti che nella "*domain ontology*" non avevano alcun legame.

### 3.2.3 Assegnazione di una distribuzione di probabilità condizionata

Devitt et al. non entrano nel merito di come definire le tabelle di probabilità condizionata dicendo che è possibile adottare uno dei sistemi già noti in letteratura.

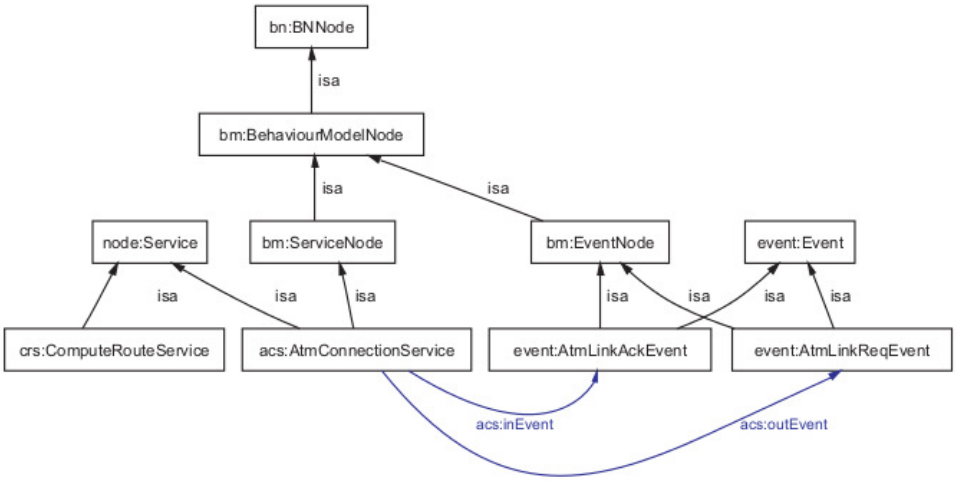


Figura 3.6: "Behaviour Model Ontology" risultante (Fonte:[4], pag.11, fig.6)

### 3.2.4 "*Behaviour Model Ontology*" risultante

La Figura 3.6 mostra il risultato dell'applicazione dei passi elencati nella Sezione 3.2 ad una ontologia che modella le prestazioni e l'affidabilità di alcune apparecchiature di telecomunicazioni.

## 3.3 BayesOWL

BayesOWL [5, 7, 6, 15] è un progetto di ricerca sviluppato fra il 2003 e il 2009 da ricercatori e professori della University of Maryland Baltimore County che ha portato notevoli contributi ai modelli di integrazione di ragionamento probabilistico nelle ontologie.

### 3.3.1 Costruzione della rete bayesiana

Il grafo della rete bayesiana corrispondente ad una ontologia espressa tramite OWL DL è costituito da:

- i nodi corrispondenti ai concetti e agli archi che rappresentano la struttura tassonomica che legano i concetti nell'ontologia (relazione "*is_a*")
- alcuni nodi logici ("*L-Nodes*"), con i relativi archi di collegamento, introdotti per modellare le relazioni logiche fra i concetti: equivalenza, unione, intersezione, disgiunzione, complemento.

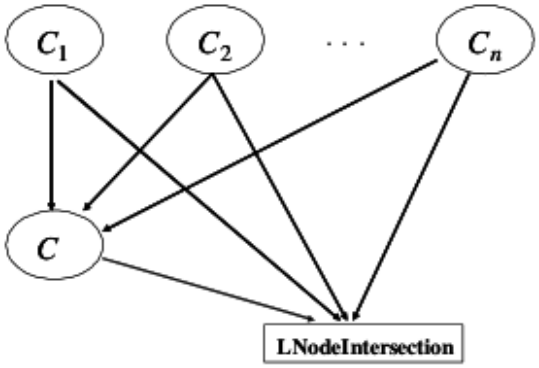


Figura 3.7: *"owl: intersectioOf"* (Fonte: [6], pag.6, fig.2)

La Figura 3.7 mostra la sotto rete risultante dalla traduzione di una classe $C$ definita come intersezione di un insieme di classi $C_n$.

Perché la rete bayesiana sia consistente con la semantica *OWL* dell'ontologia, la distribuzione di probabilità congiunta della rete bayesiana deve soddisfare il vincolo per cui tutti i nodi logici devono avere valore vero. Le tabelle di probabilità condizionata per ciascun nodo logico viene calcolata imponendo la certezza della veridicità del nodo logico in questione. Le tabelle di probabilità condizionata dei nodi concetto, sono costruite, a partire da valori definiti da esperti di dominio tramite annotazioni OWL o da valori di default (es. equiprobabilità), da un algoritmo iterativo che ricalcola i valori di probabilità in funzione dei vincoli imposti dalla necessità di dare valore vero a tutti i nodi logici (Figura 3.8).

| Male | Human | Man | |
|---|---|---|---|
| | | **True** | **False** |
| True | True | 0.5 | 0.5 |
| True | False | 0 | 1 |
| False | True | 0 | 1 |
| False | False | 0 | 1 |

| Male | Human | Man | |
|---|---|---|---|
| | | **True** | **False** |
| True | True | 0.47049 | 0.52951 |
| True | False | 0 | 1 |
| False | True | 0 | 1 |
| False | False | 0 | 1 |

Figura 3.8: Esempio di CPT iniziali e CPT ottenute dall'algoritmo iterativo (Fonte: [7], pag.7, tab.6)

### 3.3.2 Applicazioni

BayesOWL permette di applicare, in forma probabilistica, metodi di ragionamento tipici delle ontologie:

- Soddisfacibilità: l'esistenza o meno di un concetto corrispondente ad una data descrizione "$e$" può essere determinata verificando se $P(e|Q) = 0$
- Sovrapposizione: il grado di sovrapposizione o inclusione fra una descrizione "$e$" e un concetto "$C$" può essere misurato da $P(e|C,Q)$
- Sussunzione: trovare il concetto "$C$" più simile ad una descrizione "$e$" può essere calcolato con il coefficiente di similitudine di Jaccard $P(e \cap C|Q)/P(e \cup C|Q)$

Dove "Q" è la distribuzione di probabilità ottenuta rispettando il vincolo di veridicità per tutti i nodi logici.

## 3.4 Auto-extraction, representation and integration of a diabetes ontology using bayesian Networks

L'approccio descritto in [10, 9] è interessante, in quanto, gli autori generano automaticamente non solo una rete bayesiana a partire da una ontologia, ma l'ontologia di partenza stessa. La generazione automatica dell'ontologia è resa necessaria dalla constatazione che le ontologie mediche esistenti non vengono aggiornate con frequenza sufficiente a garantire la disponibilità di dati aggiornati sui recenti progressi della comunità medico-scientifica. Con una semplice analisi testuale della letteratura scientifica sul diabete di tipo II, gli autori estraggono i concetti e le relazioni fra di essi, costruendo una ontologia sulle interazioni fra geni e proteine nel diabete.
Le espressioni regolari utilizzate per l'estrazione delle informazioni necessarie alla creazione dell'ontologia permettono di catturare direttamente i concetti del dominio e le relazioni fra di essi. Esempi di relazioni catturate:

- A inibisce B
- A interagisce con B
- A attiva B

La struttura dell'ontologia viene sfruttata per costruire una rete bayesiana in modo da poter applicare ragionamento diagnostico e ragionamento causale alle manifestazioni degli effetti delle interazioni fra proteine. L'ontologia è un passaggio intermedio, fra la letteratura medica e la rete bayesiana desiderata, necessario per validare le relazioni estratte fra le proteine confrontandole con le relazioni già contenute in ontologie biomediche preesistenti (GO, BIND, KEGG).

Conoscenze a priori sulle interazioni fra entità vengono utilizzate per iniziare a strutturare la rete bayesiana definendo i primi archi fra i nodi e iniziando a popolare le tabelle di probabilità condizionata. Il resto delle relazioni e delle tabelle di probabilità condizionata vengono definite utilizzando le frequenze relative alle occorrenze delle parole chiave (concetti e relazioni).

## 3.5 Mining bayesian networks out of ontologies

L'integrazione fra reti bayesiane e ontologie proposta da Bellandi e Turini [1, 2, 3] ha la caratteristica di essere l'unico sistema noto a estrarre una rete bayesiana da una ontologia senza avere bisogno di modifiche o integrazioni alla ontologia di partenza.
Partendo da una ontologia popolata, Bellandi e Turini sviluppano un processo in tre passi che consente di "compilare" una rete bayesiana a 2 livelli, di estrarre le distribuzioni di probabilità e di eseguire ragionamento probabilistico. La rete bayesiana risultante offre capacità di ragionamento in grado di rispondere a interrogazioni probabilistiche che coinvolgono relazioni "*is_a*" e *"object property*".
Al fine di mantenere l'inferenza bayesiana computazionalmente trattabile (polinomiale rispetto al numero dei nodi) gli autori restringono il campo di applicazione del loro metodo ai "*polytree*".

### 3.5.1 Compilazione di una ontologia in una *2lBN*

#### Struttura

Per tradurre sia le relazioni fra classi ("*is_a*") che quelle fra oggetti (*"object property*") in una rete bayesiana, Bellandi e Turini utilizzano una rete a due livelli.
I nodi del livello più alto possono essere contenere una rete bayesiana: questo permette di rappresentare una gerarchia di concetti utilizzando la classe radice come nodo di livello 1 (*HLN*) al cui interno vengono rappresentate le sotto classi come rete bayesiana di livello 2 (*LLN*). Una volta imposto il

limite di trattare solamente "*polytree*" le "*object property*" vengono tradotte in archi fra nodi di primo livello (*HLR*), mentre le relazioni "*is_a*" vanno a formare gli archi fra i nodi di secondo livello (*LLR*).

**Distribuzioni di probabilità iniziali**

Supponendo che il numero di istanze appartenenti all'ontologia sia sufficientemente ampio da poterlo considerare uno spazio campione per ricavare le distribuzioni di probabilità, il processo di compilazione dell'ontologia è in grado di utilizzare le frequenze delle occorrenze delle istanze per generare le distribuzioni di probabilità iniziali.

La probabilità a priori associata ad un nodo di secondo livello è il rapporto fra il numero di istanza appartenenti a quel nodo ed il numero di istanze totali. Per ciascuna relazione fra nodi di secondo livello (*LLR*) la variabile casuale associata può assumere solamente valore vero o falso (relazione "*is_a*") con il significato di appartenenza o meno dell'istanza alla classe. Ad esempio, $P(Reseller|Company)$ esprime la probabilità che una certa istanza della classe "*Company*" sia ("*is_a*") un "*Reseller*".

| Ontology | | | | Truth values of the boolean random variables | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| D | C | Φ(C) | | | | | | | | |
| $D_1$ = Company | Company = $c_1$ | $t_1$ | | T | T | T | T | T | F | F |
| | Customer = $c_2$ | $t_2$ | | F | F | F | F | T | F | F |
| | Partnership = $c_3$ | $t_3$ | | T | T | F | F | F | F | F |
| | Jointventure= $c_4$ | $t_4$ | | F | F | F | T | F | F | F |
| | Reseller= $c_5$ | $t_5$ | | F | F | T | F | F | F | F |
| | LimitedLiabilityPS = $c_6$ | $t_6$ | | F | T | F | F | F | F | F |
| | LimitedPS= $c_7$ | $t_7$ | | T | F | F | F | F | F | F |
| | | | | | | | | | | |
| $D_2$ = Person | Person= $c_8$ | $t_8$ | | F | F | F | F | F | T | T |
| | Man = $c_9$ | $t_9$ | | F | F | F | F | F | T | F |
| | Woman = $c_{10}$ | $t_{10}$ | | F | F | F | F | F | F | T |
| Company Instances | 3648 | | | LimitedPS 345 | LimitedLiabPS 765 | Reseller 296 | Jointventure 1290 | Customer 952 | Man 1928 | Woman 976 |
| Person Instances | 2904 | | | | | | | | | |
| Ontology Instances | 6552 | | | | | | | | | |

Figura 3.9: Esempio di conteggio delle istanze (Fonte: [2], pag.53, fig.17)

Per ciascuna relazione fra nodi di primo livello (*HLR*) la variabile casuale associata è, invece, di tipo multi valore e assume tutti i valori rappresentati dagli *HLN* interessati dalla "*object property*". Ad esempio la distribuzione di probabilità condizionata per la *"object property" "hasCeo"* che coinvolge le classi *"Person"* e *"Company"* esprime la probabilità che una specifica classe di persone sia direttore generale di uno specifico tipo di azienda a seconda di come i concetti persona e azienda sono stati modellati nell'ontologia. La probabilità condizionata della "*object property*" "*hasCeo*" dipende, quindi, dal numero di istanze che popolano i nodi *LLN* degli *HLN* "*Person*" e "*Company*" che vengono coinvolti nella relazione: il calcolo della probabilità condizionata $P(Person = man|_{hasCeo} Company = reseller)$ è dato dal rapporto del numero delle istanze di persone di sesso maschile (*LLN* "man") che sono direttori generali di una azienda rivenditrice (*LLN* "*reseller*"), sul numero di direttori generali, di entrambi i sessi (*LLN* "man" $\cup$ *LLN* "woman"), di aziende rivenditrici.

### 3.5.2 Ragionamento con *2lBN*

Ragionando sui nodi *LLN* appartenenti ad uno stesso *HLN* è possibile fare inferenza su inclusione, sovrapposizione e sussunzione fra concetti (in maniera analoga a quanto visto nella 3.3.2 per BayesOWL).

Il sistema di inferenza sviluppato da Bellandi e Turini consente di ragionare anche su interrogazioni che coinvolgono le "*object property*". A tal fine gli autori definiscono un linguaggio per la formulazione e risoluzione delle interrogazioni di cui forniscono la semantica operazionale.
La definizione del linguaggio si rende necessaria perché la probabilità condizionata di un nodo *HLN* "*A*" dipende sia dai nodi evidenza che da quali archi legano il nodo "*A*" ai nodi evidenza.
Ogni arco induce un condizionamento sui nodi *HLN* incontrati durante il percorso per rispondere all'interrogazione. Infatti attraversare un arco significa far valere una "*object property*" e, di conseguenza, è necessario restringere lo spazio di probabilità del nodo destinazione dell'arco in modo che la "*object property*" sia soddisfatta.

# Capitolo 4

# Confronto

## 4.1 Prerequisiti

I metodi analizzati nel Capitolo 3 impongono come condizioni necessarie per poter essere applicati uno dei seguenti vincoli:

1. la presenza di un esperto di dominio per creare o modificare l'ontologia di partenza e per definire le distribuzioni di probabilità
2. una forma particolare dell'ontologia di partenza

Dei lavori analizzati, solamente il metodo di Bellandi e Turini (Sezione 3.5) non richiede l'intervento di un esperto di dominio per poter integrare una rete bayesiana all'ontologia di partenza. Tuttavia, il metodo proposto necessita che la struttura gerarchica dell'ontologia di partenza possa essere tradotta in un "*polytree*". Questo permette di ridurre la complessità dell'inferenza sulla rete bayesiana risultante da NP-hard a polinomiale nel numero dei nodi del grafo della rete e permette di far corrispondere la struttura dell'ontologia alla struttura di una rete bayesiana a due livelli.

OntoBayes (Sezione 3.1) ed il metodo di Devitt et al. (Sezione 3.2) necessitano della presenza di un esperto di dominio per definire: quali elementi dell'ontologia trasformare in variabili aleatorie, le relazioni di dipendenza fra le variabili e le distribuzioni di probabilità. OntoBayes è il metodo analizzato, che richiede il maggior intervento esterno perché nella rete bayesiana collega elementi che nell'ontologia non hanno legami diretti.

In BayesOWL (Sezione 3.3) un esperto di dominio deve specificare le distribuzioni di probabilità iniziali e, secondo me, dovrebbe verificare anche la validità, a livello semantico, delle distribuzioni di probabilità finali calcolate dall'algoritmo interattivo (Sezione 4.3.2).

Il metodo descritto da McGarry et al. (Sezione 3.4), pur creando ex novo una ontologia tramite una procedura di analisi testuale, invece di partire da una ontologia già esistente, richiede un esperto di dominio per poter verificare la validità dell'ontologia estratta. Tale procedimento di verifica viene compiuto confrontando i concetti e le relazioni estratti dalla letteratura scientifica con quelli già modellati nelle ontologie mediche esistenti.

## 4.2 Parte dell'ontologia modellata dalla rete bayesiana

Implicitamente, o esplicitamente (BayesOWL, Bellandi e Turini), tutti i metodi si limitano ad un sotto insieme proprio delle caratteristiche di una ontologia OWL DL. La trattazione di una ontologia OWL-Full propria (cioè non esprimibile in OWL DL) comporterebbe la necessità di dover trattare:

- la non decidibilità della logica sottostante l'ontologia
- la possibile presenza di cicli nelle definizioni di classi e istanze (in OWL-Full le classi possono essere trattate come istanze)
- la non disgiunzione fra "*object property*" e "*datatype property*"

Considerando le difficoltà sopra elencate e che, la quasi totalità dei sistemi software per modellare le ontologie e dei motori di ragionamento supporta OWL DL, ritengo che la scelta di concentrare gli sforzi di ricerca su OWL DL sia una scelta obbligata.

Sfortunatamente, però, nessuno dei lavori è in grado di modellare in una rete bayesiana tutti i componenti dell'ontologia di partenza: tutti i lavori analizzati restringono le tipologie di elementi dell'ontologia che vengono modellati nella rete bayesiana. La presenza di queste restrizioni è dovuta alla mancanza di un modello di traduzione dell'intera capacità espressiva dell'ontologia in una rete bayesiana (Sezione 4.3.1).

La rete bayesiana ottenuta tramite l'uso di OntoBayes cattura "*object property*" e "*datatype property*".

Devitt et al. affrontano la trattazione delle "*object property*" , "*datatype property*" e delle relazioni "*is_a*".

Gli autori di BayesOWL dichiarano espressamente di considerare "la struttura tassonomica dell'ontologia", quindi relazioni "*is_a*" e gli operatori logici rappresentati dai nodi logici (Sezione 3.3.1).

Il metodo di Bellandi e Turini tratta "*object property*" e relazioni "*is_a*". È l'unico metodo che utilizza la *ABox* dell'ontologia nella modellazione della rete bayesiana.

McGarry et al. non specificano cosa il loro metodo sia in grado di trattare, ma dalle tipologie di relazioni fra geni estratte dalla letteratura sul diabete si evince che vogliano modellare "*object property*".

## 4.3 Traduzione da ontologia a rete bayesiana

La definizione e descrizione del processo di traduzione da ontologia a rete bayesiana costituisce una parte importante del lavoro svolto dagli autori degli articoli analizzati; tanto che in 2 casi (OntoBayes Sezione 3.1e Devitt et al. Sezione 3.2) l'intero sforzo di ricerca è stato rivolto, quasi unicamente, alla definizione del processo di traduzione. La particolare attenzione riservata a questo aspetto dell'integrazione fra ontologie e reti bayesiane è giustificabile dalle considerazioni che: non è ancora stato definito un metodo di traduzione che permetta di trasferire tutti gli elementi modellati da una ontologia in una rete bayesiana, che la ricerca in questo campo è relativamente recente.

Una proprietà desiderabile del processo di traduzione è quella di mantenere, quando possibile, inalterata la struttura dell'ontologia di partenza, consentendo così di poter utilizzare ragionamento ontologico e ragionamento bayesiano sulle medesime entità.
Dei processi di traduzione proposti, solamente quelli di McGarry et al. (Sezione 3.4)[1] e Bellandi e Turini (Sezione 3.5) non modificano l'ontologia di partenza. Gli altri processi di traduzione richiedono l'aggiunta di annotazioni o classi di concetti tramite estensioni di OWL che sono definite ad hoc per ciascun lavoro e, pertanto, difficilmente integrabili fra di loro. Devitt et al. (Sezione 3.2) realizzano un approccio ibrido: un'ontologia ausiliaria viene utilizzata congiuntamente con l'ontologia di partenza per definire la rete bayesiana. Anche con questo metodo rimane l'impossibilità di utilizzare ragionamento bayesiano e ragionamento semantico sulle medesime entità dato che l'onto-

[1] una volta ottenuta l'ontologia dall'analisi testuale della letteratura

logia ausiliaria ridefinisce le relazioni fra concetti e filtra i valori delle "*datatype property*" tradotti nella rete bayesiana.

### 4.3.1 Parte strutturale

La prima parte del processo di traduzione dell'ontologia consiste nella creazione della parte strutturale della rete bayesiana, cioè la definizione dei nodi e degli archi di un grafo orientato aciclico.

Come specificato nella parte introduttiva del Capitolo 1, sia le ontologie che le reti bayesiane usano una struttura a grafo per modellare i concetti rappresentati e le relazioni fra di essi. Nonostante questa struttura comune, i lavori analizzati propongono metodologie differenti per la traduzione del grafo di una ontologia nel grafo di una rete bayesiana. Infatti, come analizzato nella Sezione 4.2 ciascuno dei metodi proposti trasforma nella rete bayesiana parti differenti del grafo dell'ontologia, senza riuscire a modellare per intero il potere espressivo (e la relativa complessità) dell'ontologia.

In OntoBayes i nodi della rete bayesiana rappresentano *"object property"* e *"datatype property*". Dal momento che, nell'ontologia, non esiste un collegamento diretto fra *"object property*" e *"datatype property"* in grado di definire come una proprietà possa influenzarne[2] un'altra, gli autori definiscono una estensione di OWL che consente di annotare esplicitamente la relazione di dipendenza fra due proprietà. Ad esempio, in questo modo, è possibile esprimere il legame di dipendenza fra il prezzo di un prodotto ("*datatype property*") e la probabilità che un cliente sia intenzionato a comprarlo ("*object property*" fra cliente e prodotto) (Figura 3.1). Il processo di traduzione è in realtà uno schema in grado di guidare il procedimento manuale di un esperto di dominio: ogni qual volta sia necessario eseguire un ragionamento probabilistico, l'esperto di dominio dovrà scegliere quali proprietà dell'ontologia faranno parte della conoscenza da modellare e dovrà annotare le relazioni di dipendenza fra di esse.

Devitt et al. propongono un metodo di traduzione che non modifica l'ontologia di partenza e che permette, ad un esperto di dominio, di specificare tutte le informazioni necessarie alla creazione della rete bayesiana dentro il sistema di strumenti software utilizzati per la modellazione e il ragionamento ontologici.

Detta *O* l'ontologia di partenza, il metodo richiede la definizione di una prima ontologia ausiliaria *A,* in cui specificare le classi di *O* (relazioni "is_a") da tradurre in nodi della rete bayesiana definendo in *A* una partizione delle classi concetto di *O* fra "concetti di interesse" e "concetti non di interesse" (Figura 3.4). Le ontologie *O* e *A* vengono combinate in una nuova ontologia *B* che permette di

[2] nel senso bayesiano del termine

specificare restrizioni sui valori delle "*datatype property*" (Figura 3.5). L'ontologia *B* così ottenuta è la base di partenza per ottenere la struttura della rete bayesiana: una volta definito un sistema di regole ad hoc per la creazione degli archi rappresentati dalle "*object property*" è possibile applicare il sistema di ragionamento semantico su *B* per ottenere un'ontologia che rappresenti la struttura della rete bayesiana.
Il metodo proposto è adatto solamente ad esperti di ontologie, infatti, per la traduzione delle tre componenti dell'ontologia modellate: "*is_a*", "*datatype property*", "*object property*" è necessario definire due ontologie ed un sistema di regole applicabili dal motore di ragionamento semantico.

La traduzione della parte strutturale definita in BayesOWL non richiede interventi esterni e si basa solamente sulle informazioni contenute nell'ontologia di partenza. Le classi dei concetti e le relazioni "is_a" vengono tradotte, rispettivamente, in nodi della rete bayesiana e archi diretti dalla classe genitore alla sottoclasse figlio. La traduzione degli operatori logici richiede che una procedura automatica definisca nuovi nodi nella rete bayesiana e crea gli archi diretti dai concetti coinvolti nelle operazioni logiche verso il nodo che rappresenta l'operatore logico(Figura 3.7).

Anche il metodo descritto da Bellandi e Turini non richiede alcun intervento esterno per poter tradurre una ontologia in una rete bayesiana e si basa unicamente sulle informazioni contenute nell'ontologia di partenza. A differenza del metodo proposto in BayesOWL, Bellandi e Turini utilizzano una rete bayesiana a 2 livelli. Come visto nella Sezione 4.1 l'ontologia di partenza deve essere rappresentabile tramite un "*polytree*". La struttura del "*polytree*" garantisce che le relazioni che modellano le *"object property"* sussistano solamente fra le classi radice di una gerarchia, cioè quelle classi che hanno come super classe solamente la classe top "*owl:Thing*". Di conseguenza, tutte le altre classi, saranno coinvolte solamente in relazione "*is_a*". Questa partizione fra classi e relazioni viene tradotta in una rete bayesiana a due livelli: i nodi di primo livello rappresentano le radici delle gerarchie e contengono al loro interno i nodi di secondo livello, che rappresentano le sottoclassi della gerarchia di appartenenza. Per mantenere la struttura del "*polytree*" non è possibile definire archi che coinvolgano nodi a livelli differenti, quindi gli archi fra nodi di primo livello rappresentano le "*object_property*", mentre gli archi fra nodi di secondo livello rappresentano relazioni "*is_a*".

McGarry et al. partono dall'analisi della letteratura per estrarre un'ontologia prima di tradurre l'informazione in una rete bayesiana. La conoscenza estratta potrebbe essere modellata direttamente in una rete bayesiana saltando il processo di estrazione dell'ontologia, ma in questo modo, non sarebbe possibile verificare la correttezza delle informazioni estratte dall'analisi testuale. Per il tipo di pattern utilizzati nell'analisi testuale, l'ontologia estratta è rappresentabile come insieme di triple

RDF (soggetto, verbo, predicato) e contiene solamente "*object property*". La traduzione da questo tipo di ontologia ad una rete bayesiana è banalmente la trasposizione di classi in nodi e relazioni in archi.

### 4.3.2 Parte probabilistica

Dei metodi descritti solamente quelli di Bellandi e Turini e McGarry et al. calcolano automaticamente le distribuzioni di probabilità associate alla rete bayesiana.

McGarry et al. usano il conteggio della frequenze relative delle parole chiave trovate negli articoli di ricerca analizzati per costruire le tabelle di probabilità condizionata.

Bellandi e Turini richiedono che l'ontologia di partenza sia popolata e che le istanze possano essere considerate uno spazio campione statisticamente significativo in modo da poter definire le tabelle di probabilità condizionata contando le frequenze delle istanze coinvolte nelle varie relazioni. Nel caso in cui fosse necessario utilizzare una ontologia non sufficientemente popolata, un esperto di dominio può fornire dei "data set" da utilizzare per l'apprendimento delle probabilità o direttamente le tabelle di probabilità condizionata. Alcune semplici interrogazioni probabilistiche che coinvolgono concetti con legami noti ad un esperto di dominio (es. il fatto che due classi siano disgiunte) possono essere un primo test per capire se le distribuzioni di probabilità ricavate dall'ontologia sono semanticamente significative.

In OntoBayes le distribuzioni di probabilità devono essere completamente specificate all'interno dell'ontologia tramite le estensioni OWL sviluppate dagli autori. Le classi OWL appositamente create per descrivere le distribuzioni di probabilità a priori e quelle condizionate sono sufficienti per descrivere le distribuzioni di proprietà, ma richiedono la definizione di una classe per ogni valore di probabilità presente nelle tabelle; considerando che le variabili aleatorie in OntoBayes sono di tipo multi valore è facile prevedere un proliferare di classi aggiunte all'ontologia di partenza al solo scopo di modellare i valori di proprietà.

Devitt et al. non entrano nel merito di come specificare le distribuzioni di probabilità, limitandosi a dichiarare che la conoscenza di certe implicazioni in relazioni "*is_a*" possono essere inserite tramite la scrittura di regole apposite, analoghe a quelle utilizzate per la creazione degli archi, durante la compilazione della parte strutturale. Un'altra possibilità per inserire informazioni probabilistiche è rappresentata dai nodi della "*Behaviour Model Ontology*" che vengono utilizzati per restringere valori sulle "*datatype property*" e potrebbero specificare distribuzioni di probabilità a priori.

In BayesOWL affinché la rete bayesiana, ottenuta tramite la compilazione della parte strutturale,

mantenga la semantica dell'ontologia di partenza deve essere imposto sulle tabelle di probabilità condizionata il vincolo che tutti i nodi logici devono avere valore vero. Il processo di traduzione della parte probabilistica della rete bayesiana è composto da tre passi: nel primo un esperto di dominio inserisce dei vincoli sulle probabilità a priori e su alcune probabilità condizionate, nel secondo viene assegnato un valore di default in tutti quei casi in cui una tabella non sia stata completamente definita, nell'ultimo passo l'esecuzione di un algoritmo iterativo modifica le tabelle di probabilità condizionata in modo che i vincoli imposti dalle tabelle di probabilità dei nodi logici siano rispettati. Il procedimento così definito presenta dei problemi per ciascuno dei passi descritti:

- l'algoritmo iterativo non è in grado di garantire un risultato semanticamente corretto nel caso emergano conflitti fra i vincoli imposti dai nodi logici e quelli forniti dall'esperto di dominio sui valori di probabilità
- l'esperto di dominio non può specificare alcun vincolo per le tabelle di probabilità condizionata in cui siano presenti più di una variabile evidenza[3]
- l'assegnazione dei valori di default, indipendentemente dalla distribuzione utilizzata, porta ad inconsistenze semantiche nella definizione delle tabelle di probabilità condizionata di concetti disgiunti creati dalla combinazione di più evidenze

Un esempio che mostra le conseguenze degli ultimi due difetti elencati è riportato in Figura 3.8: la probabilità che un essere umano di sesso maschile sia un uomo è più bassa di quella che non lo sia. Nell'ontologia la definizione di "*Man*" era data semplicemente da "*Man,is_a,Male*" e "*Man,is_a,Human*", ma l'esperto di dominio non ha potuto specificare quello che sarebbe stato il giusto vincolo semantico $P(Man|Male, Human) = 1$ e la distribuzione dei valori di default ha assegnato l'equiprobabilità ai due valori della variabile booleana generando una distribuzione di probabilità condizionata che non rispetta la semantica della classe "*Man*".

## 4.4 Tipologie di ragionamento probabilistico

Le differenze presenti fra i vari lavori, che sono state analizzate nelle Sezioni 4.1, 4.2, 4.3 di que-

[3] questo vincolo viene specificato, sia negli articoli che descrivono BayesOWL, che sul sito web dell'applicazione omonima, ma non viene data alcuna spiegazione del motivo di questa limitazione.

sto capitolo, traggono origine da / sono causa di differenti tipologie di ragionamento probabilistico implementate dagli autori dei lavori analizzati.

I metodi generici, cioè creati come strumenti applicabili ad un ampio spettro di problemi e domini, sono BayesOWL e quello di Bellandi e Turini. Entrambi i metodi offrono una forma probabilistica dei ragionamenti ontologici di: sussunzione di concetti, appartenenza ad una classe, sovrapposizioni fra classi.

In BayesOWL questo permette, non solo di utilizzare forme di ragionamento semantico in presenza di dati incerti e incompleti, ma anche di ottenere risposte più accurate e precise di quelle che potrebbero venir fornite da molti sistemi di ragionamento semantico su descrizioni non probabilistiche. Ad esempio in [6] gli autori dimostrano come la ricerca del concetto più simile alla descrizione $\neg Male \sqcap Animal$ sull'insieme di classi {Animal, Male, Female, Human, Man, Woman} dia come risultato "*Female*" mentre il sistema di ragionamento Racer[4] proponga "*Animal*" come soluzione.

Nel metodo sviluppato da Bellandi e Turini le forme di ragionamento semantico, elencate sopra, devono essere ristrette ai soli nodi *LLN* appartenenti ad uno stesso *HLN* a causa della separazione imposta dal "*polytree*" ai nodi *LLN* appartenenti ad *HLN* distinti. Nei casi d'uso reali, la restrizione non dovrebbe essere così forte come appare: dato che la rete bayesiana traduce fedelmente la struttura dell'ontologia è sempre possibile eseguire i ragionamenti ontologici prima sull'ontologia e poi, ripetere il ragionamento in forma probabilistica all'interno del nodo *HLN* corrispondente alla gerarchia identificata come risposta dal ragionamento sull'ontologia. Inoltre il concetto di esistenza e sussunzione spesso sono definiti all'interno di una stessa gerarchia (come nell'esempio sopra citato da [6]) per non fornire risultati troppo generici per essere realmente utili.
Le forme di ragionamento semantico riguardano solamente le relazioni "*is_a*" e i rispettivi nodi *LLN*, ma il motore di ragionamento probabilistico sviluppato da Bellandi e Turini è in grado di rispondere ad interrogazioni probabilistiche che coinvolgono un numero arbitrario di "*object property*" e i rispettivi nodi *HLN* e *LLN*. Ad esempio in [1] gli autori mostrano come combinando le tre forme di inferenza su rete bayesiana (Sezione 2.2.3) sia possibile rispondere alla seguente interrogazione: "*Which is the probability that a Patent project is led by person which is CEO of a company operating in the financial sector?*". L'interrogazione viene formulata tramite il linguaggio definito da Bellandi e Turini e il motore di ragionamento la risolve tramite l'applicazione delle scomposizioni delle probabilità condizionate in prodotti di probabilità di eventi indipendenti, l'applicazione ricorsiva dei metodi di inferenza bayesiana e il condizionamento dello spazio di probabilità (come analizzato

[4] http://www.sts.tu-harburg.de/~r.f.moeller/racer/

nella Sezione 3.5.2).

Gli altri lavori analizzati sviluppano un'integrazione fra ontologie e reti bayesiane al fine di poter eseguire ragionamento probabilistico in domini di applicazione ben definiti:

- OntoBayes è stato sviluppato come approccio per ragionare su incertezza e probabilità in un sistema di supporto alle decisioni ("decision making") da applicare a prodotti assicurativi: la gestione del rischio in caso di disastri naturali
- Il sistema di Devitt et al. nasce con il duplice scopo di modellare la conoscenza su una rete di telecomunicazioni adattiva e auto-configurante (conoscenza in continuo mutamento, spesso incompleta e quindi difficilmente modellabile con una ontologia tradizionale) e di fornire i risultati del ragionamento probabilistico direttamente come input ai dispositivi di monitoraggio e configurazione della rete di telecomunicazioni per influenzarne la configurazione
- Il sistema di McGarry et al. viene creato appositamente per poter studiare gli ultimi risultati presenti in letteratura sullo studio del diabete di tipo II, partendo dalla costatazione che le ontologie mediche esistenti non sono vengono aggiornate sufficientemente in fretta da permettere ai medici e ai biologi di poterle utilizzare come fonti di informazioni per i loro esperimenti

Questi sistemi "*specializzati*" non offrono sistemi di ragionamento avanzati come quelli di Bellandi e Turini o BayesOWL, ma si limitano ad applicare i processi di inferenza bayesiana sulle reti costruite.

## 4.5 Considerazioni finali

I lavori analizzati e confrontati nel corso dei Capitoli 3 e 4 sono stati scelti in quanto rappresentativi dello stato dell'arte dei metodi di integrazione fra ontologie e reti bayesiane. I cinque lavori sono ben distinti ed eterogenei: nessuno è in grado di modellare tutte le caratteristiche di un'ontologia in una rete bayesiana ed ogni lavoro si concentra su un sottoinsieme diverso di caratteristiche, fornendo un metodo di traduzione ad hoc per il sottoinsieme scelto.

Tre (OntoBayes, Devitt et al., McGarry et al.) dei cinque lavori analizzati sfruttano il potere di inferenza delle rete bayesiane per poter risolvere un problema specifico, mentre i rimanenti due

(BayesOWL, Bellandi e Turini) utilizzano l'inferenza della rete bayesiana come mezzo per costruire forme di ragionamento più articolate e complesse.

Il fatto che gli autori di tre lavori ritengano che l'inferenza bayesiana sia sufficiente per migliorare le possibilità di ragionamento in tre domini completamente differenti fra loro — la medicina, la valutazione del rischio in campo assicurativo, le reti di telecomunicazioni auto-configuranti — deve essere visto, secondo me, come un segnale positivo sulla validità delle scelta di utilizzare le reti bayesiane come complemento probabilistico per le ontologie.

Bellandi, Turini e gli autori di BayesOWL mostrano come sia possibile costruire sistemi di ragionamento in grado di rispondere ad interrogazioni complesse ed espressive, utilizzando l'inferenza bayesiana come fondamenta di un lavoro di ricerca maggiormente collegato alle aree di ricerca tradizionali dell'informatica (algoritmi, data mining, semantica dei linguaggi, logica, ecc).

# Capitolo 5

# Conclusioni

Per gli esperti di dominio è necessario poter integrare nelle ontologie la trattazione di conoscenza incompleta o incerta. Come specificato nella Sezione 1.1 e analizzato in dettaglio nel Capitolo 4, non esiste un modello unico di integrazione fra reti bayesiane e ontologie.

In questa tesi sono stati analizzati e confrontati cinque diversi modelli di integrazione scelti in quanto rappresentativi dello stato dell'arte. Sono stati confrontati caratteristiche tecniche, punti di forza e di debolezza di ciascun metodo evidenziando e tenendo conto delle differenti motivazioni pratiche che hanno spinto gli autori a sviluppare i vari metodi. In particolare il confronto ha messo in risalto le differenze nel diverso grado di automazione nell'integrazione, le conoscenze ritenute necessarie perché un esperto di dominio possa applicare un determinato modello, le parti dell'ontologia prese in considerazione, le tecniche di traduzione da ontologia a rete bayesiana e le tipologie di ragionamento probabilistico.

Dei metodi legati alla risoluzione di un particolare problema, ho cercato di fornire una analisi "generale", astraendo, quando possibile, dallo specifico contesto di applicazione al fine di permettere, così, un confronto delle caratteristiche dei modelli svincolato dalle particolari restrizioni imposte dal dominio applicativo.

Questa tesi è il primo lavoro di analisi e confronto sui modelli di integrazione di reti bayesiane e ontologie di cui sia a conoscenza. La tesi presenta lo stato dell'arte nei modelli di integrazione di reti bayesiane e ontologie e offre un confronto di ampio raggio, infatti i cinque modelli analizzati sono distinti ed eterogenei e rappresentano bene le tipologie di integrazioni descritte nella letteratura presa in considerazione. La tesi mostra, anche, le difficoltà da affrontare e le scelte da ponderare per tradurre alcune parti, piuttosto che altre, da un'ontologia in una rete bayesiana.

Questo lavoro può, dunque, essere utile come base di partenza per ulteriori ricerche sulla letteratura, ma anche per lo sviluppo di nuove tecniche di integrazione fra reti bayesiane ed ontologie in quanto analizza le difficoltà da affrontare e mostra alcuni limiti intrinseci della traduzione da ontologia a rete bayesiana.

# Bibliografia


[1] Bellandi Andrea and Turini Franco. Extending ontology queries with bayesian network reasoning. In *Proceedings of the IEEE 13th international conference on Intelligent Engineering Systems*, INES'09, pages 149–154, Piscataway, NJ, USA, 2009. IEEE Press.

[2] Andrea Bellandi. *Extending Ontology Queries with Bayesian Network Reasoning*. PhD thesis, IMT Institute for Advanced Studies, Lucca, 2008.

[3] Turini Franco Bellandi Andrea. Mining bayesian networks out of ontologies. *JIIS*, -:–, 2011. submitted.

[4] ANN DEVITT, B. Danev, and K. Matusikova. Constructing bayesian networks automatically using ontologies. In *Proceedings of Second Workshop on Formal Ontologies Meets Industry*. FOMI, 2006.

[5] Z Ding and Y Peng. A probabilistic extension to ontology language owl. In *in HICSS 04: Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS 04) - Track 4*, 2004.

[6] Z Ding, Y Peng, and Pan. R.: Bayesowl: Uncertainty modeling in semantic web ontologies. In *Soft Computing in Ontologies and Semantic Web. Studies in Fuzziness and Soft Computing*. Springer, 2006.

[7] Zhongli Ding, Yun Peng, and Rong Pan. A bayesian approach to uncertainty modeling in owl ontology. In *Proceedings of the International Conference on Advances in Intelligent Systems Theory and Applications*, Luxembourg, November 2004. Has three GS keys: UMdqaFCbMakJ, M2aqATowR4cJ, GTtycwLF6z0J.

[8] Tom Gruber. Ontology, the encyclopedia of database systems. In *Encyclopedia of Database Systems*, pages 1963–1965. Springer-Verlag, 2009.

[9] Ken McGarry, Sheila Garfield, and Stefan Wermter. Auto-extraction, representation and integration of a diabetes ontology using bayesian networks. *Computer-Based Medical Systems, IEEE Symposium on*, 0:612–617, 2007.

[10] Kenneth McGarry, Sheila Garfield, and Stefan Wermter. Auto-extraction, representation and integration of a diabetes ontology using bayesian networks. In *CBMS*, pages 612–617, 2007.

[11] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[12] W3C. Owl web ontology language reference.

[13] Y Yang and J Calmet. Ontobayes approach to corporate knowledge. In *In Proceeding of the I6th International Symposium on Methodologies for Intelligent Systems (ISMIS 06), LNCS/LNAI*. Springer, 2006.

[14] Yi Yang and Jacques Calmet. Ontobayes: An ontology-driven uncertainty model. In *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce Vol-1 (CIMCA-IAWTIC'06) - Volume 01*, CIMCA '05, pages 457–463, Washington, DC, USA, 2005. IEEE Computer Society.

[15] Shenyong Zhang, Yun Peng, and Xiaopu Wang. Bayesowl: A prototype system for uncertainty in semantic web. In *Proceedings of the 2009 International Conference on Artificial Intelligence,*, pages 678–684, July 2009.