Learning Rules on Top of Ontologies: An Inductive Logic Programming Approach

Francesca A. Lisi

lisi@di.uniba.it

LACAM group

Dipartimento di Informatica

Università degli Studi di Bari

Via Orabona, 4 - 70126 Bari - Italy

CILC 2006 @ Dipartimento di Informatica – Università degli Studi di Bari

26 June 2006

Overview

- **Motivation**
- Background
- The general framework

- Two instantiations of the general framework
- Conclusions

Motivation:
Ontology engineering

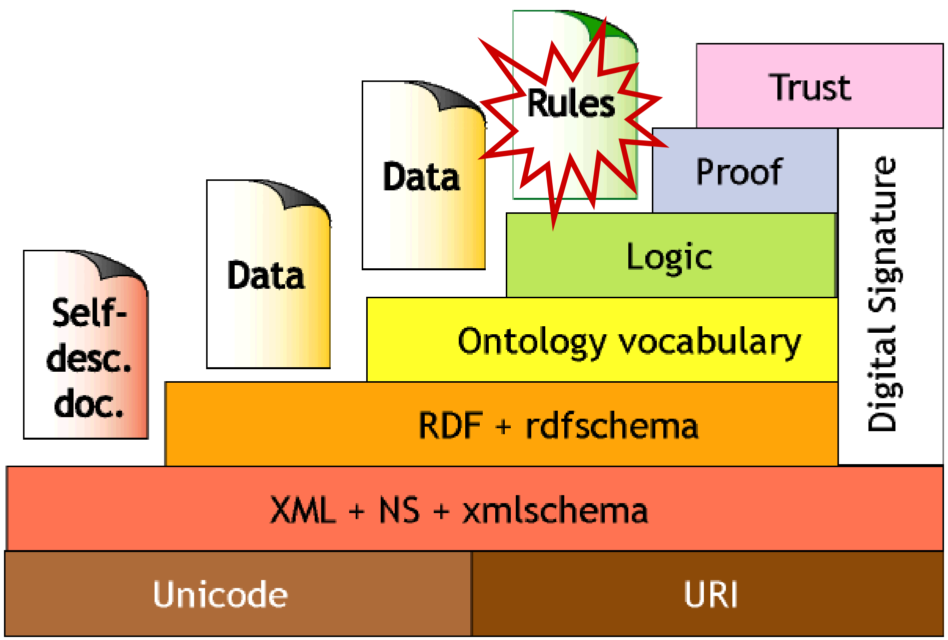
- An **ontology** is a formal explicit specification of a shared conceptualization for a domain of interest (Gruber, 1993)
- Description logics are the most used formalisms for specifying ontologies, but ..
- ..Some concepts are better specified by means of **rules**!!

Building rules is a demanding task

Rule induction

Motivation:
The Semantic Web

SWRL (Semantic Web Rule Language)
- submitted to W3C for standardization
- integration of OWL and RuleML
- undecidable ☹



- Decidable fragment of SWRL?
- $\mathcal{AL}$-log (Donini et al., 1998)!

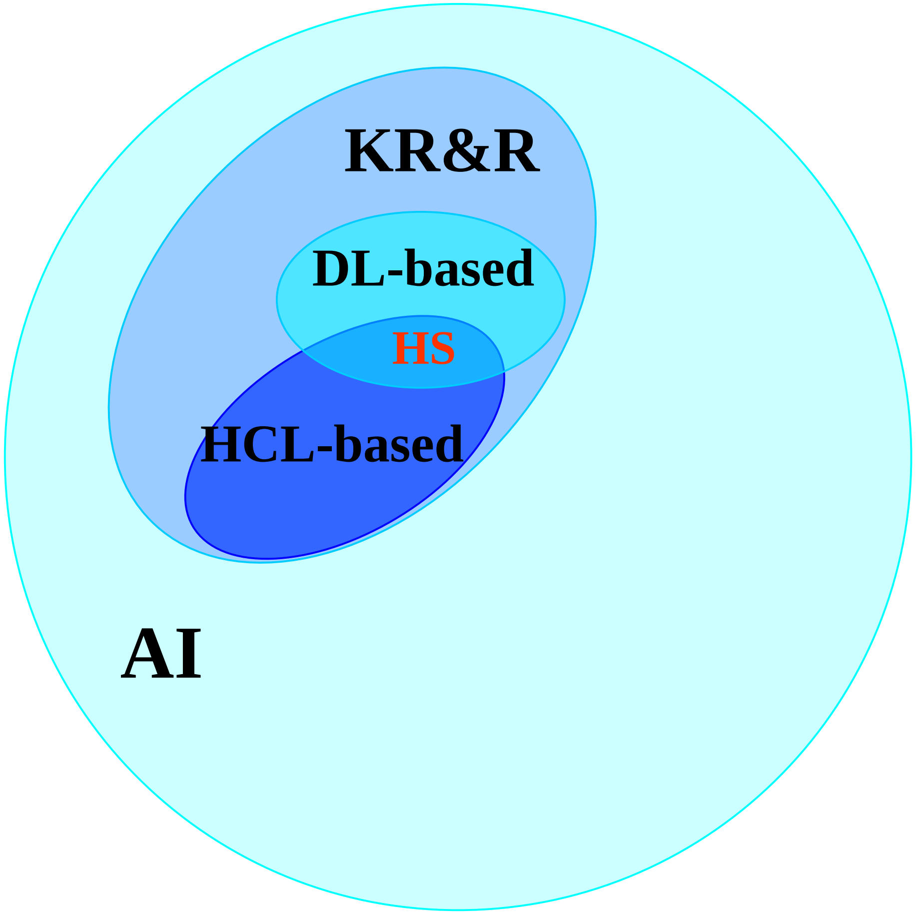
- Induction for logics?
- Inductive Logic Programming!

Overview

- Motivation
- **Background**
  - The hybrid system $\mathcal{AL}$-log
  - Inductive Logic Programming
- The general framework

- Two instantiations of the general framework
- Conclusions

DL-Horn Hybrid KR&R Systems



- Horn Clausal Logic (HCL) and Description Logics (DLs) are fragments of first-order logic
- They can not be compared wrt expressive power but can be combined to obtain more expressive KR&R systems
  - $\mathcal{AL}$-log = $\mathcal{ALC}$ + Datalog (Donini et al., 1998)
  - CARIN (Levy & Rousset, 1998)

The hybrid system $\mathcal{AL}$-log: syntax

$$\mathcal{B} = < \Sigma , \Pi >$$

equality axioms $C \equiv D$
inclusion axioms $C \sqsubseteq D$
concept assertions $a{:}C$
role assertions $<a, b>{:}R$

constrained Datalog clauses
$\alpha_0 \leftarrow \alpha_1, \ldots, \alpha_m \;\&\; \gamma_1, \ldots, \gamma_n$
where γ_j must be
a $\mathcal{ALC}$ concept assertion

- Only one Datalog literal in the head
- Only positive Datalog literals in the body
- Constraints <u>must</u> refer to variables occurring in the Datalog part
- Variables in the Datalog part <u>can</u> be constrained

The hybrid system $\mathcal{AL}$-log: semantics

$$\mathcal{J} = (\mathcal{I}_O, \mathcal{I}_H)$$

interpretation for Σ

Herbrand interpretation for Π_D

- $\mathcal{J}$ satisfies $\mathcal{B}$ iff
  - it satisfies Σ, and
  - for each clause $\alpha_0 \leftarrow \alpha_1, \ldots, \alpha_m \;\&\; \gamma_1, \ldots, \gamma_n$, for each of its ground instances $\alpha'_0 \leftarrow \alpha'_1, \ldots, \alpha'_m \;\&\; \gamma'_1, \ldots, \gamma'_n$, either there exists one γ'_i, $1 \leq i \leq n$, that is not satisfied by $\mathcal{J}$ or $\alpha'_0 \leftarrow \alpha'_1, \ldots, \alpha'_m$ is satisfied by $\mathcal{J}$
- OWA of $\mathcal{ALC}$ and CWA of Datalog do not interfere (safeness)
- UNA holds for $\mathcal{ALC}$ and *ground* Datalog

The hybrid system $\mathcal{AL}$-log: reasoning

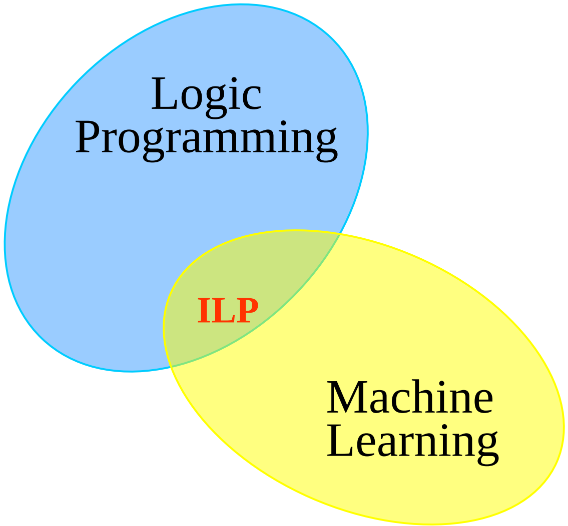
Consistency of Σ

- Tableau calculus
  - instance checks (a:C wrt Σ?)
  - subsumption checks ($C \subseteq D$ wrt Σ?)

Query answering wrt Π

- constrained SLD-resolution
  - decidable
  - complete by refutation
    - It tries to obtain the empty constrained Datalog clause (= only constraints) by applying SLD-resolution, then
    - It verifies that constraints in the empty constrained Datalog clause are consistent wrt Σ by applying tableau calculus

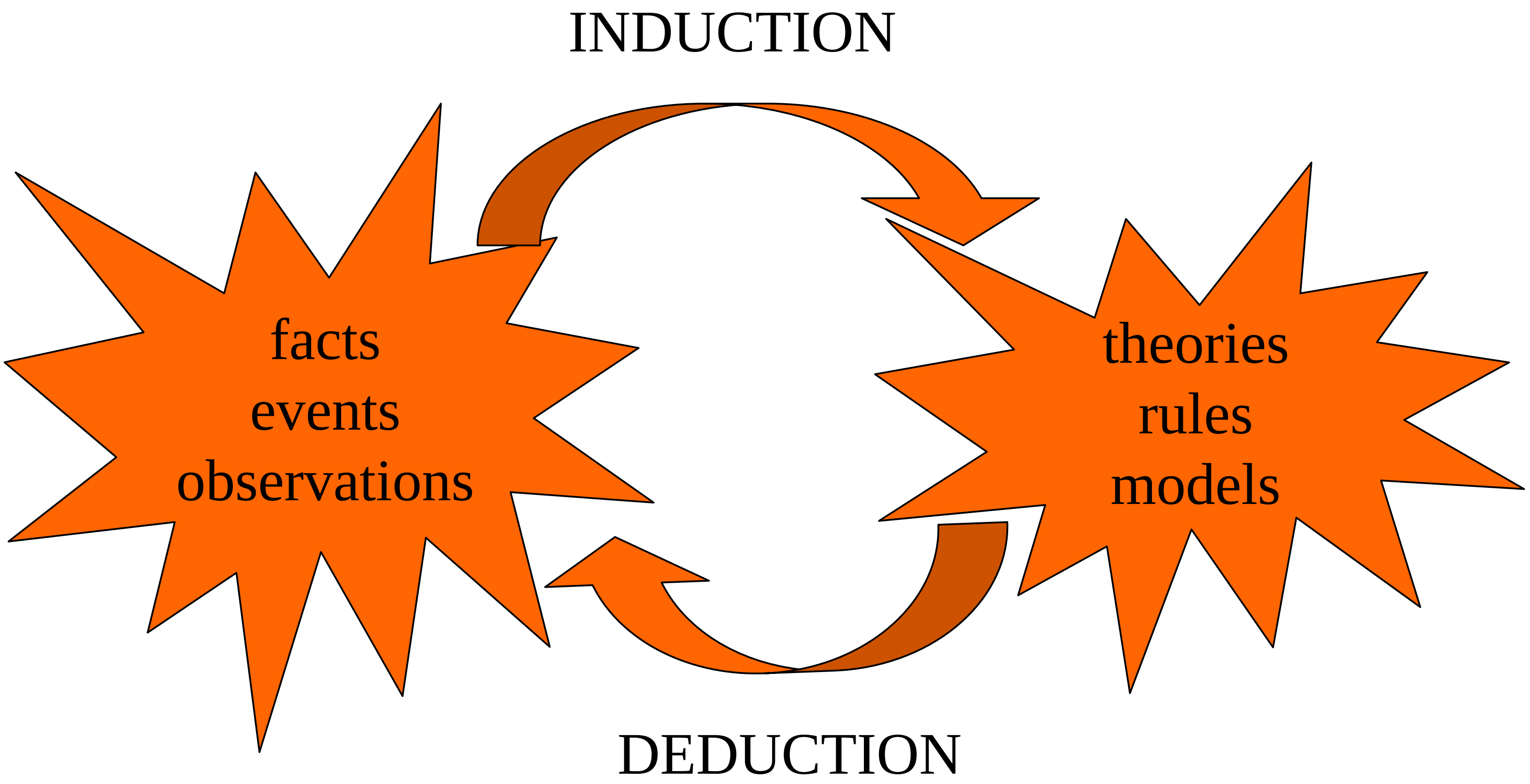
Inductive Logic Programming



- *Originally* Induction of rules from observations and background knowledge within the representation framework of Horn clausal logic (Muggleton, 1990)
  - scope of induction: discrimination
  - task: prediction

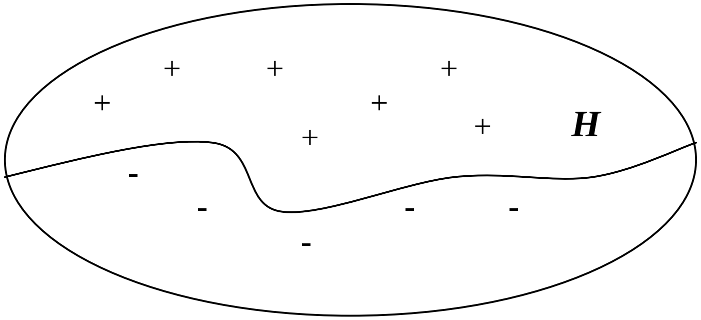
- *Currently* Induction of rules from observations and background knowledge within the representation framework of first-order logic (fragments)
  - scope of induction: discrimination/characterization
  - task: prediction/description

Inductive Learning:
induction vs deduction

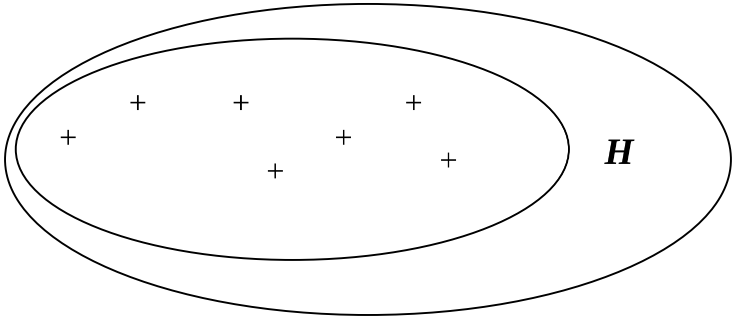


Inductive Learning:
Scope of induction

- Prediction



- Description



Overview

- Motivation
- Background
- **The general framework**
  - Problem statement
  - The language of hypotheses
  - The generality order
  - The coverage relations
- Two instantiations of the general framework
- Conclusions

The general framework: problem statement

Goal

Induction of rules from

- a set $\mathcal{O} = \{o_i\}$ of observations, and
- a background knowledge $\mathcal{K}$, $\mathcal{K} \cap \mathcal{O} = \emptyset$,

within the representation framework of $\mathcal{AL}$-log

General = independent from the scope of induction

The general framework: The language of hypotheses

Hypotheses as constrained Datalog clauses compliant with:

- linkedness
  - $h(X) \leftarrow b(X,Y) \;\&\; X{:}C,\ Y{:}D$ linked
  - $h(X) \leftarrow b(X,Y),\ c(Z) \;\&\; X{:}C,\ Y{:}D$ not linked

- connectedness
  - $h(X) \leftarrow b(X,Y) \;\&\; X{:}C,\ Y{:}D$ connected
  - $h(Z) \leftarrow b(X,Y) \;\&\; X{:}C,\ Y{:}D$ not connected

- Object Identity (OI)
  - *In a formula terms denoted with different symbols represent different entities of the domain* (Semeraro et al., 1998)
  - Extension of the Unique Names Assumption from the semantics of $\mathcal{ALC}$ to the syntax of $\mathcal{AL}$-log -> OI-substitution

The general framework: The generality order (1/2)

$\mathcal{B}$-subsumption: A model-theoretic definition

- $\mathcal{B}=\mathcal{K}\cup\mathcal{O}$
- $H_1, H_2 \in \mathcal{L}$

H_1 $\mathcal{B}$-subsumes H_2 if
- for every model $\mathcal{J}$ of $\mathcal{B}$ and
- every ground atom α such that H_2 covers α under $\mathcal{J}$,

we have that H_1 covers α under $\mathcal{J}$.

H_1 is at least as general as H_2 under $\mathcal{B}$-subsumption, $H_1 \geq_{\mathcal{B}} H_2$, iff H_1 $\mathcal{B}$-subsumes H_2

The general framework: The generality order (2/2)

$\mathcal{B}$-subsumption: A proof-theoretic definition

- $\mathcal{B}=\mathcal{K}\cup\mathcal{O}$
- $H_1, H_2 \in \mathcal{L}$
- σ a Skolem substitution for H_2 w.r.t. $\{H_1\}\cup\mathcal{B}$

$H_1 \geq_{\mathcal{B}} H_2$ iff there exists a substitution θ for H_1 such that

- $head(H_1)\theta = head(H_2)$
- $\mathcal{B}\cup body(H_2)\sigma \mid\!- body(H_1)\theta\sigma$
- $body(H_1)\theta\sigma$ is ground.

- Checking $\geq_{\mathcal{B}}$ in $\mathcal{AL}$-log is decidable.

The general framework: The coverage relations (1/2)

Learning from implications

- $H = q(\mathbf{X}) \leftarrow body(H) \in \mathcal{L}$
- o_i is a ground constrained Datalog clause with head $q(\mathbf{a}_i)$

$H \in \mathcal{L}$ covers $o_i \in O$ under entailment w.r.t. $\mathcal{K}$ iff

- $\mathcal{K} \cup H \models o_i$
- $\mathcal{K} \cup body(o_i) \cup H \vdash q(\mathbf{a}_i)$

The general framework: The coverage relations (2/2)

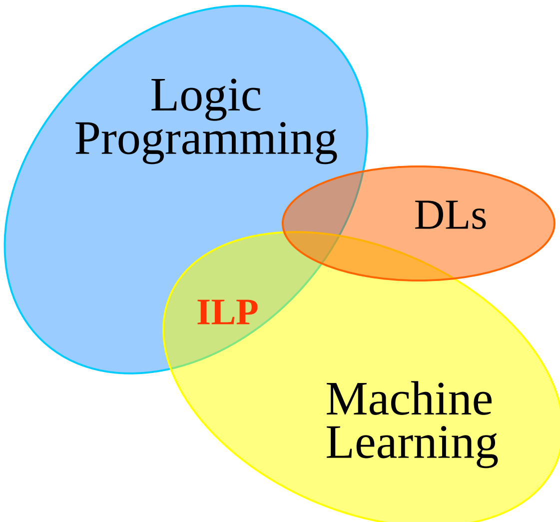
Learning from interpretations

- $H = q(\boldsymbol{X}) \leftarrow body(H) \in \mathcal{L}$
- $o_i = (q(\boldsymbol{a_i}), \mathcal{A}_i)$ where $\mathcal{A}_i$ is a set of ground Datalog facts

$H \in \mathcal{L}$ covers $o_i \in O$ under interpretations w.r.t. $\mathcal{K}$ iff

- $\mathcal{K} \cup \mathcal{A}_i \cup H \models q(\boldsymbol{a_i})$
- $\mathcal{K} \cup \mathcal{A}_i \cup H \vdash q(\boldsymbol{a_i})$

The general framework: Conclusions



Related work on learning in hybrid languages

- Frisch, *Sorted downward refinement: Building background knowledge into a refinement operator for ILP*, ILP'99
- Rouveirol and Ventos, *Towards learning in CARIN-$\mathcal{ALN}$*, ILP'00
- Kietz, *Learnability of description logic programs*, ILP'02

Why $\mathcal{AL}$-log is to be preferred to CARIN (Levy & Rousset, 1998)

- Safe interaction between DL and HCL part
- decidable reasoning mechanisms
- expressive and deductive power *enough* for the *actual* needs of the Semantic Web

Overview

- Motivation
- Background
- The general framework

- **Instantiations of the framework**
  - The setting
  - Task 1: Frequent pattern discovery
  - Task 2: Conceptual clustering
- Conclusions

The setting

Characteristic induction from interpretations

- $\mathcal{L}$ a language of hypotheses
- $\mathcal{K}$ a background knowledge
- M($\mathcal{B}$) model constructed from $\mathcal{B}=\mathcal{K}\cup O$
- $o_i = (q(\boldsymbol{a_i}), \mathcal{A}_i)$ where $\mathcal{A}_i$ is a set of ground Datalog facts

Finding $\mathcal{H}\subseteq\mathcal{L}$ such that

- $\mathcal{H}$ is true in M($\mathcal{B}$) and
- for each $H\in\mathcal{L}$, if H is true in M($\mathcal{B}$) then $\mathcal{H} \models H$

Frequent Pattern Discovery
(at multiple levels of description granularity)

Given

- a taxonomic ontology Σ
- a reference concept $C_{ref} \in \Sigma$
- some task-relevant concepts C_{tsk}'s from Σ
- a data source Π
- a multi-grained language $\mathcal{L}=\{\mathcal{L}^l\}_{1\leq l \leq maxG}$ of patterns

Find the set of all patterns expressible with $\mathcal{L}$ that describe associations between C_{ref} and C_{tsk}'s and are frequent in **r=** $\Sigma \cup \Pi$

Frequent Pattern Discovery
(at multiple levels of description granularity

The definition of *frequency*

Given (also)

- a set $\{minsup^l\}_{1 \leq l \leq maxG}$ of support thresholds
- an evaluation function *supp* for patterns

A pattern **P** $\in \mathcal{L}^l$ with *supp*(P, **r**)=s is frequent in **r** iff

- s $\geq$ $minsup^l$
- all ancestors of *P* w.r.t. Σ are frequent in **r**

A pattern Q$\in$ $\mathcal{L}^h$, $h<l$, is an ancestor of P iff it can be obtained from P by replacing each concept C occurring in P with a concept D$\in$ $\mathcal{T}^h$ such that C is a sub-concept of D

Frequent Pattern Discovery
(at multiple levels of description granularity)

Find in:

- the on-line CIA World Fact Book (*data set* **CIA**)

frequent patterns describing:

- Middle East countries (*reference concept* MiddleEastCountry)

with respect to:

- the religions believed (*task-relevant concepts* from the hierarchy rooted in Religion)
  - e.g., the Muslim religion is a monotheistic religion
- the languages spoken (*task-relevant concepts* from the hierarchy rooted in Language)
  - e.g., the Indo-Iranian language is an Indo-European language

at three levels of description granularity (maxG=3)

Instantiating the framework

The data set **r**

- A knowledge base in $\mathcal{AL}$-log

$$\mathcal{B} = < \Sigma , \Pi >$$

An example

- MiddleEastCountry $\equiv$ AsianCountry $\cap \exists$Hosts.MiddleEasternEthnicGroup
- AsianCountry $\subseteq$ Country
- ‘IR’:AsianCountry, <‘IR’,’Arab’>:Hosts, ’Arab’:MiddleEasternEthnicGroup

- believes(CountryID, ReligionN) $\leftarrow$ religion(CountryID, ReligionN,Perc)
  & CountryID:Country, ReligionN: Religion
- religion(‘IR’, ‘Shia’, 89)

Instantiating the framework

The language $\mathcal{L}$ of patterns

- Unary conjunctive queries in $\mathcal{AL}$-log called $\mathcal{O}$-queries

$$Q = q(X) \leftarrow \alpha_1, \ldots, \alpha_m \;\&\; X{:}C_{ref}, \gamma_1, \ldots, \gamma_n$$

distinguished variable

reference concept

- Generated starting from:
  - a set $\mathcal{A}$ of Datalog predicate names
  - a set Γ' of $\mathcal{ALC}$ concept names (belonging to the level l of description granularity)

Instantiating the framework

The language $\mathcal{L}$ of patterns: an example

- $\mathcal{A}$ = {believes/2, speaks/2}
- C_{ref} = MiddleEastCountry
- Γ^1 = {Religion,Language}
- Γ^2 = {MonotheisticReligion, …, IndoEuropeanLanguage}
- Γ^3 = {MuslimReligion, …, IndoIranianLanguage}

Q_0 = q(X) $\leftarrow$ & X: MiddleEastCountry   trivial $\mathcal{O}$-query

Q_1 = q(X) $\leftarrow$ believes(X,Y) & X:MiddleEastCountry, Y:Religion $\in \mathcal{L}^1$

Q_2 = q(X) $\leftarrow$ believes(X,Y), speaks(X,Z) &
X:MiddleEastCountry, Y:MonotheisticReligion, Z:IndoEuropeanLanguage $\in \mathcal{L}^2$

Q_3 = q(X) $\leftarrow$ believes(X,Y), speaks(X,Z) &
X:MiddleEastCountry, Y:MuslimReligion, Z:IndoIranianLanguage $\in \mathcal{L}^3$

Instantiating the framework

The evaluation function *supp* for $\mathcal{L}$

- It computes the percentage of individuals of C_{ref} that satisfy the $\mathcal{O}$-query
- It is based on the coverage relation under interpretations

$$\mathbf{supp(Q, \mathcal{B})} = |\, \mathrm{answerset}(Q, \mathcal{B}) \,| \;/\; |\, \mathrm{answerset}(Q_t, \mathcal{B}) \,|$$

Instantiating the framework

The evaluation function *supp* for $\mathcal{L}$: an example

Q = q(X) ← believes(X,Y) & X:MiddleEastCountry, Y:JewishReligion
answerset(Q, $\mathcal{B}$) = {'IR', 'IL', 'SYR'}

Q_t = q(X) ← & X: MiddleEastCountry
|answerset(Q_t, $\mathcal{B}$)| = 15

supp(Q, $\mathcal{B}$) = 20%

20% Middle East countries believe a Jewish religion

Instantiating the framework

The generality order

- $\geq_{\mathcal{B}}$ is monotone w.r.t support
  - if $P \geq_{\mathcal{B}} Q$ then $supp(P, \mathcal{B}) \geq supp(Q, \mathcal{B})$

- $\geq_{\mathcal{B}}$ satisfies **all** the assumptions underlying Mannila's levelwise search method!
  - If a generality order $\geq$ over $\mathcal{L}$ can be defined such that $\geq$ is monotone w.r.t. *supp*, the lattice $(\mathcal{L}, \geq)$ can be searched with a breadth-first strategy level by level of depth (Mannila & Toivonen, 1997)
- The **ILP system $\mathcal{AL}$-QuIn** implements this method for $\mathcal{L}$ being a language of $\geq_{\mathcal{B}}$–ordered $\mathcal{O}$-queries

Instantiating the framework

The generality order: examples

Q_1 = q(X) $\leftarrow$ believes(X,Y) & X:MiddleEastCountry, Y:Religion

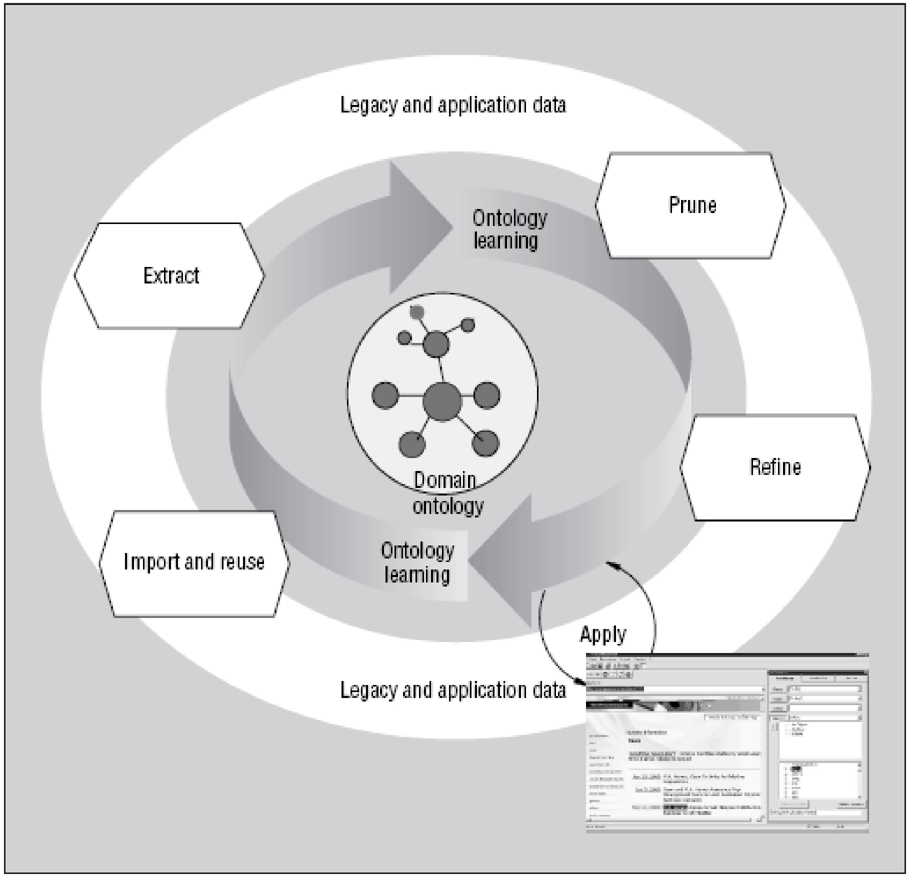
Q_5 = q(X) $\leftarrow$ believes(X,Y) & X:MiddleEastCountry, Y:MonotheisticReligion

Q_6 = q(X) $\leftarrow$ believes(X,Y), believes(X,Z) &
X:MiddleEastCountry, Y:Religion

$Q_1 \geq_{\mathcal{B}} Q_5$ (but not viceversa due to constrained SLD-resolution)

$Q_1 \geq_{\mathcal{B}} Q_6$ (but not viceversa due to the OI bias)

Ontology Refinement



- Adaptation of an existing ontology to a specific domain or the needs of a particular user (Maedche & Staab, 2001)
  - Adding new relations
  - Adding new concepts

Concept Refinement

Given

- a taxonomic ontology Σ
- a data source Π
- a concept $C_{ref} \in \Sigma$ (*reference concept*)
- a language $\mathcal{L}$

Find

a directed acyclic graph (DAG) $\mathcal{G}$ of concepts C_i such that:

1. $int(C_i) \in \mathcal{L}$
2. $ext(C_i) \subset ext(C_{ref})$

Concept Refinement (reformulated)

- Concept Refinement ≅ Conceptual Clustering
  - Form of unsupervised learning that aims at determining <u>not only</u> the clusters <u>but also</u> their descriptions in some formalism

Conceptual Clustering: A pattern-based approach (1/2)

- **Assumption:** frequent patterns as clues of data clusters

- **Two-phased method**:
  1. detect *emerging* concepts (known ext, unknown int)
  2. turn emerging concepts into *fully-specified* ones (known ext, known int)

It can rely on an algorithm for frequent pattern discovery, e.g. $\mathcal{AL}$-QuIn!

Conceptual Clustering: A pattern-based approach (2/2)

Different frequent patterns can have the same answer set!

P = q(A) ← speaks(A,B), believes(A,C) &

A:MiddleEastCountry, B:AfroAsiaticLanguage, C:MonotheisticReligion

answerset(P, $\mathcal{B}$) = {'IR', 'SA'}

ext(C)= {'IR', 'SA'}

Q = q(A) ← speaks(A,B), believes(A,C) &

A:MiddleEastCountry, B:ArabicLanguage, C:MuslimReligion

answerset(Q, $\mathcal{B}$)=answerset(P, $\mathcal{B}$)

int(C)=?

Conceptual Clustering: The choice criterion (1/2)

Language bias + **Search bias**

Based on $\mathcal{O}$-query form, e.g.

- Minimum level of description granularity
- Nro ontologically constrained variables

Based on $\mathcal{B}$-subsumption, e.g.

- Most general description
- Most specific description

Conceptual Clustering: The choice criterion (2/2)

Examples

Descriptions must have all the variables ontologically constrained by concepts from the 2nd granularity level on

m.g.d.

int(C) =

q(A) ← speaks(A,B), believes(A,C) &

A:MiddleEastCountry,

B:AfroAsiaticLanguage,

C:MonotheisticReligion

ext(C)= {'IR', 'SA'}

m.s.d.

int(C) =

q(A) ← speaks(A,B), believes(A,C) &

A:MiddleEastCountry,

B:ArabicLanguage,

C:MuslimReligion

ext(C)= {'IR', 'SA'}

Instantiations of the framework: related work

Frequent pattern discovery

- (Han & Fu, 1999)
  - It <u>can</u> deal with hierarchies
  - It <u>doesn't</u> adopt FOL
- (Dehaspe & Toivonen, 1999)
  - It <u>can't</u> deal with hierarchies
  - It <u>does</u> adopt FOL (Datalog)

Conceptual clustering

- Very few works!
  - <u>Some</u> of them <u>do</u> adopt FOL
  - <u>None</u> of them <u>do</u> adopt a pattern-based approach
- (Vrain, 1996)
  - It applies a top-down incremental but distance-based method
  - It <u>does</u> adopt FOL (an object-logical representation)

Bibliography (1/2)

- F.A. Lisi & F. Esposito (2006). *Two Orthogonal Biases for Choosing the Intensions of Emerging Concepts in Ontology Refinement*. To appear in: ECAI-06 proceedings.
- F.A. Lisi (2006). *A Pattern-based Approach to Conceptual Clustering in FOL*. In H. Schärfe, P. Hitzler, P. Øhrstrøm (Eds.), Conceptual Structures: Inspiration and Application (ICCS-06), LNAI 4068, Springer: Berlin.
- F.A. Lisi (2006). *Practice of Inductive Reasoning on the Semantic Web.* To appear in: J.J. Alferes, J. Bailey, W. May, U. Schwertel (Eds.), Principles and Practice of Semantic Web Reasoning 2006, LNCS ?, Springer: Berlin.
- F.A. Lisi (2005). *Principles of Inductive Reasoning on the Semantic Web*. In F. Fages and S. Soliman (Eds.), Principles and Practice of Semantic Web Reasoning, LNCS 3703, 118-132, Springer: Berlin.
- F.A. Lisi & F. Esposito (2005). *ILP meets Knowledge Engineering: A Case Study*. In S. Kramer and B. Pfahringer (Eds.), *Inductive Logic Programming*, LNAI 3625, 209-226, Springer: Berlin.

Bibliography (2/2)

- F.A. Lisi & D. Malerba (2004). *Inducing Multi-Level Association Rules from Multiple Relations*. Machine Learning, 55:175-210.
- F.A. Lisi & F. Esposito (2004). *Efficient Evaluation of Candidate Hypotheses in $\mathcal{AL}$-log*. In R. Camacho, R. King, and A. Srinivasan (Eds.), Inductive Logic Programming, LNAI 3194, 216-233, Springer: Berlin.
- F.A. Lisi & D. Malerba (2003). *Ideal Refinement of Descriptions in $\mathcal{AL}$-log*. In T. Horvath and A. Yamamoto (Eds.), Inductive Logic Programming, LNAI 2835, 215-232, Springer: Berlin.
- F.A. Lisi & D. Malerba (2003). *Bridging the Gap between Horn Clausal Logic and Description Logics in Inductive Learning*. In A. Cappelli and F. Turini (Eds.), AI*IA 2003: Advances in Artificial Intelligence, LNAI 2829, 53-64, Springer: Berlin.
- F.A. Lisi (2002). *An ILP setting for Object-Relational Data Mining*. PhD Thesis, University of Bari, Dept. of CS.

Future work

- Study of learnability (i.e. computational complexity of learning) of $\mathcal{AL}$-log
- Extension of the framework towards more expressive hybrid languages (closer to SWRL)
- Instantiation of the framework into cases of discriminant induction (predictive rules)
- More implementation work, e.g. post-processing from $\mathcal{AL}$-log to SWRL

Future work on ResearchCyc

- ResearchCyc is the most comprehensive outcome of the Cyc project, began by Douglas Lenat in 1984 and currently carried out by Cycorp, Inc. (Austin, TX,U.S.A.)
- The Cyc project's objective was to codify, in machine-usable form, the millions of pieces of knowledge that comprise human common sense.
  - "People have silly reasons why computers don't really think. The answer is we haven't programmed them right; they just don't have much common sense. There's been only one large project to do something about that, that's the famous Cyc project.“ - Marvin Minsky, MIT, May 2001
- ”Refining the ResearchCyc ontology with Inductive Logic Programming” is **winner of The 2006 Cyc Prize Competition** for the best research proposal!

Thanks for
your attention!