Seminar aus Maschinellem Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

Seminar im Wintersemester 2011/2012

Frederik Janssen, Heiko Paulheim

Fachgebiet Knowledge Engineering

Organisatorisches

TECHNISCHE UNIVERSITÄT DARMSTADT

- Zeitslot:
  - Dienstag, 16:15 – 17:55
- Infos zur Veranstaltung:
  - http://www.ke.tu-darmstadt.de/lehre/ws-11-12/seminar-aus-maschinellem-lernen
- Wichtig: bitte in TUCaN anmelden!

Anforderungen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Thema aussuchen und erarbeiten
- 1-2 Papers werden vorgegeben
  - zusätzliche eigene Recherchen wünschenswert
- Vortrag im Seminar halten
  - ca. 30 Minuten plus Fragen und Diskussion
  - Vorträge und Themen können gern vorab besprochen werden
    - Sprechstunde Frederik Janssen: Do 11-12, Heiko Paulheim Di 15-16
    - oder Termin vereinbaren

- Regelmäßige *aktive* Teilnahme (max. 3 mal fehlen)
  - aktiv: Fragen stellen, mitdiskutieren
  - merke: geht in die Bewertung mit ein,
    dafür aber keine Ausarbeitung nötig

Themenauswahl

TECHNISCHE UNIVERSITÄT DARMSTADT

- Themen werden gleich (grob) vorgestellt
- Papers sind alle auf der Webseite verlinkt; diese Folien werden ebenfalls online gestellt

- Bis Donnerstag abend eine Mail schicken
  - Themenwünsche (drei mit Prioritäten)
  - Datumswunsch:
    - eher früh (November)
    - eher mittelfrüh (Dezember bis Anfang Januar)
    - eher spät (Mitte Januar bis Semesterende)

- Zeitplan wird voraussichtlich am Freitag veröffentlicht

Inhalte/Ziele

TECHNISCHE UNIVERSITÄT DARMSTADT

- Schwerpunkte dieses Semester:
  - Regellernen
  - Maschinelles Lernen im Semantic Web
- Ziele:
  - Überblick in die Gebiete geben
  - Einblicke in aktuelle Forschung geben
  - Synergie zwischen Regellernen und Semantic Web herstellen

Regellernen

TECHNISCHE UNIVERSITÄT DARMSTADT

- eine der ältesten Disziplinen im Bereich des Maschinellen Lernens (in etwa in den 1960er Jahren erfunden)
- *daher:*
  - sehr gut untersucht
  - eine Vielzahl von starken Algorithmen
  - extensiv genutzt
- *aber:*
  - noch immer sind gewisse Komponenten des Regellernens nicht gut verstanden, u.a. Heuristiken oder Suchalgorithmen
- es gibt verschiedene Teilbereiche des Regellernens, die 2 wichtigsten sind Propositional Rule Learning (Aussagenlogisch) und Relational Rule Learning (Inductive Logic Programmming)
  - in diesem Seminar beschäftigen wir uns mit beiden Themenbereichen

Propositionales und relationales Regellernen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Propositional Rule Learning
- *gegeben:*
  - eine Menge von Trainingsbeispielen
- *gesucht:*
  - eine Regelmenge, wobei die Regeln direkt auf der Trainingsmenge arbeiten

- Relational Rule Learning
- *gegeben:*
  - eine Menge von Trainingsbeispielen und eine Menge von Hintergrundwissen
- *gesucht:*
  - eine Regelmenge, wobei die Regeln aus dem Hintergrundwissen gebildet werden

Regellern-Algorithmen
auszugsweise

TECHNISCHE UNIVERSITÄT DARMSTADT

- Propositional Rule Learning
  - AQ (Michalski, 1969)
  - CN2 (Clark and Niblett, 1989)
  - RIPPER (Cohen, 1995)
  - PART (Frank and Witten, 1998)
- Relational Rule Learning
  - FOIL (Quinlan, 1990)
  - Progol (Muggleton, 1995)

Separate-and-conquer Regellernen

TECHNISCHE UNIVERSITÄT DARMSTADT

- auch als Covering-Strategie bezeichnet
- meist genutzte Strategie im propositionalen Regellernen
- guter Überblicksartikel:
  - Separate-and-conquer Rule Learning (Fürnkranz, 1999)
- Ablauf:
  - lerne eine Regel, die einen Teil der Daten beschreibt
  - füge diese Regel der Regelmenge hinzu
  - entferne alle Beispiele, die von der Regel abgedeckt sind
  - lerne die nächste Regel (solange bis alle positiven Beispiele abgedeckt sind)

Regellernen: Themen
General Topic

TECHNISCHE UNIVERSITÄT DARMSTADT

- paper: *Fast Counting with AV-Space for Efficient Rule Induction*
  - AV-spaces werden eingeführt
  - um Kandidatenregeln zu bewerten muss die Abdeckung dieser gezählt werden, was einem Durchlauf durch alle Beispiele entspricht
  - AV-spaces sind eine Datenstruktur, die diesen Durchlauf unnötig macht und damit das Lernen von Regeln effizienter macht
  - die Struktur ist für Separate-and-conquer Algorithmen geeignet

Regellernen: Themen
Probabilistic Rule Learning & Ranking

TECHNISCHE UNIVERSITÄT DARMSTADT

- paper: *Probabilistic Rule Learning*
  - der Lerner FOIL wird um probabilistische Methoden erweitert
- paper: *Learning to rank cases with classification rules*
  - üblicherweise werden Entscheidungsbäume genutzt um zu ranken (Beispiele werden nach der Sicherheit der Vorhersage gerankt)
  - im paper wird das Ranken mit Regeln beschrieben (3 verschiedene Ansätze werden disktuiert)
  - abschließend: empirische Studie

Regellernen: Themen
Features and Rule Learning Algorithms

TECHNISCHE UNIVERSITÄT DARMSTADT

- paper: *Explicit Feature Construction and Manipulation for Covering Rule Learning Algorithms*
  - eine Methode zur Generierung von Features wird eingeführt
  - theoretische Fundierung, dass Feature-Generierung und Regelkonstruktion getrennt betrachtet werden sollten
  - Betrachtungen zu fehlenden Werten und ungenauen Attributwerten (bei numerischen Features)

Regellernen: Themen
User-based Evaluation of Rules

TECHNISCHE UNIVERSITÄT DARMSTADT

- papers: *Gray Box Robustness Testing of Rule Systems* & *Declarative Specification and Interpretation of Rule-Based Systems*
  - Erweiterung von DATALOG
  - neue Evaluaierungsstrategien
  - Betrachtungen zu fehlerhaften Benutzereingaben
  - Fallstudie

Regellernen: Themen
Search algorithms for Rule Learners

TECHNISCHE UNIVERSITÄT DARMSTADT

- paper: *Finding a short and accurate decision rule in disjunctive normal form by exhaustive search*
  - Einführung des Algorithmus EXPLORE (Regellerner, der eine vollständige Suche benutzt)
  - Pruning-Strategie ähnlich wie in RIPPER (durch Validationsdatenset)
  - Vergleich zu 8 anderen Regellernern
  - interessante Aussage:
    - je besser/tiefer die Suche desto besser die Genauigkeit der gefundenen Regelmengen
    - kontrovers zu vorher entdecktem Problem des „Oversearching"
    - Literatur:
      - Oversearching and layered search in empirical learning (Quinlan and Cameron-Jones, 1995)
      - A Re-evaluation of the Over-Searching Phenomenon in Inductive Rule Learning (Janssen and Fürnkranz, 2009)

Regellernen: Themen
Statistics

TECHNISCHE UNIVERSITÄT DARMSTADT

- paper: *A Statistical Approach to Rule Learning*
  - 2 Hauptprobleme beim Regellernen:
    - Berechnungsaufwand
    - Überanpassung
  - statistischer Ansatz mit gewichteten Regeln; formuliert als konvexes Optimierungsproblem
    - finde Regeln mit großem Abstand und kleiner Varianz
  - Überanpassung wird durch Selektion von Regelmengen verhindert
- paper: *Efficient Rule Ensemble Learning using Hierarchical Kernels*
  - Regularisierung hierarchischer Kernels kann als Bias zu kleinen Regeln interpretiert werden
  - Reformulierung der Untermengenselektion, um Optimierungsproblem konvex zu machen

Regellernen: Themen
Verification/Anomalies of ontologies with rules

TECHNISCHE UNIVERSITÄT DARMSTADT

- papers: *Verification and Refactoring of Ontologies With Rules* & *Towards the Verification of Ontologies with Rules* & *Anomalies in ontologies with rules*
  - papers überlappen stark
  - Ontologien mit Regeln
    - Probleme:
      - Evaluierung muss angepasst werden
      - Anomalien treten auf (z.B. Zyklen oder Redundanz)
    - erste Lösungsansätze werden gegeben

Semantic Web in a Nutshell

TECHNISCHE UNIVERSITÄT DARMSTADT

- Das "klassische" Web

aus Sicht des Menschen:

```
<html>
  …
  <b>Dr. Mark Smith</b>
  <i>Physician</i>
  Main St. 14
  Smalltown
  Mon-Fri 9-11 am
  Wed 3-6 pm
  …
</html>
```

aus Sicht des Computers:

Dr. Mark Smith
Physician
Main St. 14
Smalltown
Mon-Fri 9-11 am
Wed 3-6 pm

Print in bold: „hmf298hmmhudsa“
Print in italics: „mj2i9ji0“
Print normal: „fdsah
02hfadsh0um2m0adsmf0ihm
asdfjköfdsa298ndsfmij32mio
lk2mjpoimjiofdpmsajiomjm“

Semantic Web in a Nutshell

TECHNISCHE UNIVERSITÄT DARMSTADT

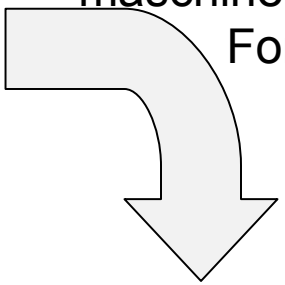
- Problem: Suche ist sehr schlecht
- Information von verschiedenen Seiten muss manuell kombiniert werden

- Lösungsansatz Web Mining:
  - Der Maschine beibringen, im klassischen Web zu lesen
- Lösungsansatz Semantic Web:
  - Information für Maschinen aufbereiten

Semantic Web in a Nutshell

TECHNISCHE UNIVERSITÄT DARMSTADT

```
<html>
  …
  <b>Dr. Mark Smith</b>
  <i>Physician</i>
  Main St. 14
  Smalltown
  Mon-Fri 9-11 am
  Wed 3-6 pm
  …
</html>
```

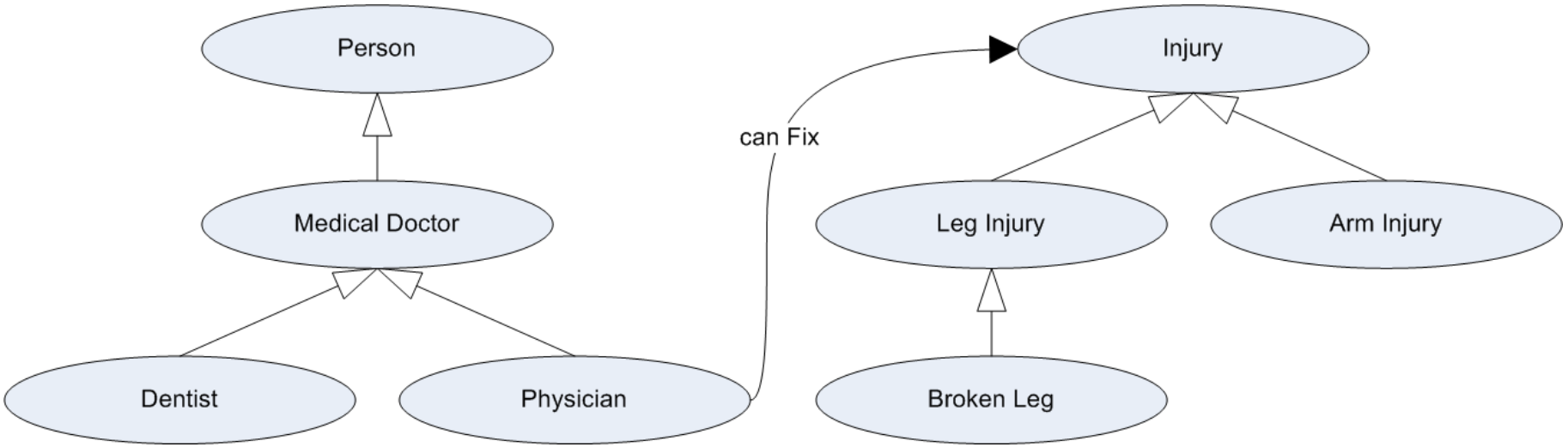



```
PREFIX …

ms:person a dbpedia:Physician.
ms:person address ms:addr.
ms:person name „Mark Smith“
ms:addr street „Main St. 14“
ms:addr city „Smalltown“
…
```

Semantic Web in a Nutshell

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ontologien: formalisiertes Wissen über eine Domäne



Semantic Web in a Nutshell

TECHNISCHE UNIVERSITÄT DARMSTADT

- Aus Instanzdaten (A-box) und Ontologie (T-box) kann jetzt komplexe Information abgeleitet werden

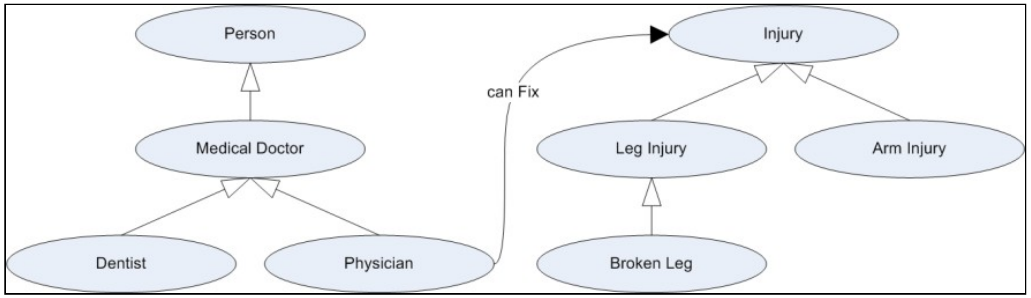
```
SELECT ?name WHERE {
       ?p a Physician,
       ?p address ?a
       ?a city „Smalltown“ }
```

```
SELECT ?name WHERE {
       ?p a MedicalDoctor, … }
```

-
```
SELECT ?name WHERE {
       ?p a Person,
       ?p canFix ?bl,
       ?bl a BrokenLeg }
```

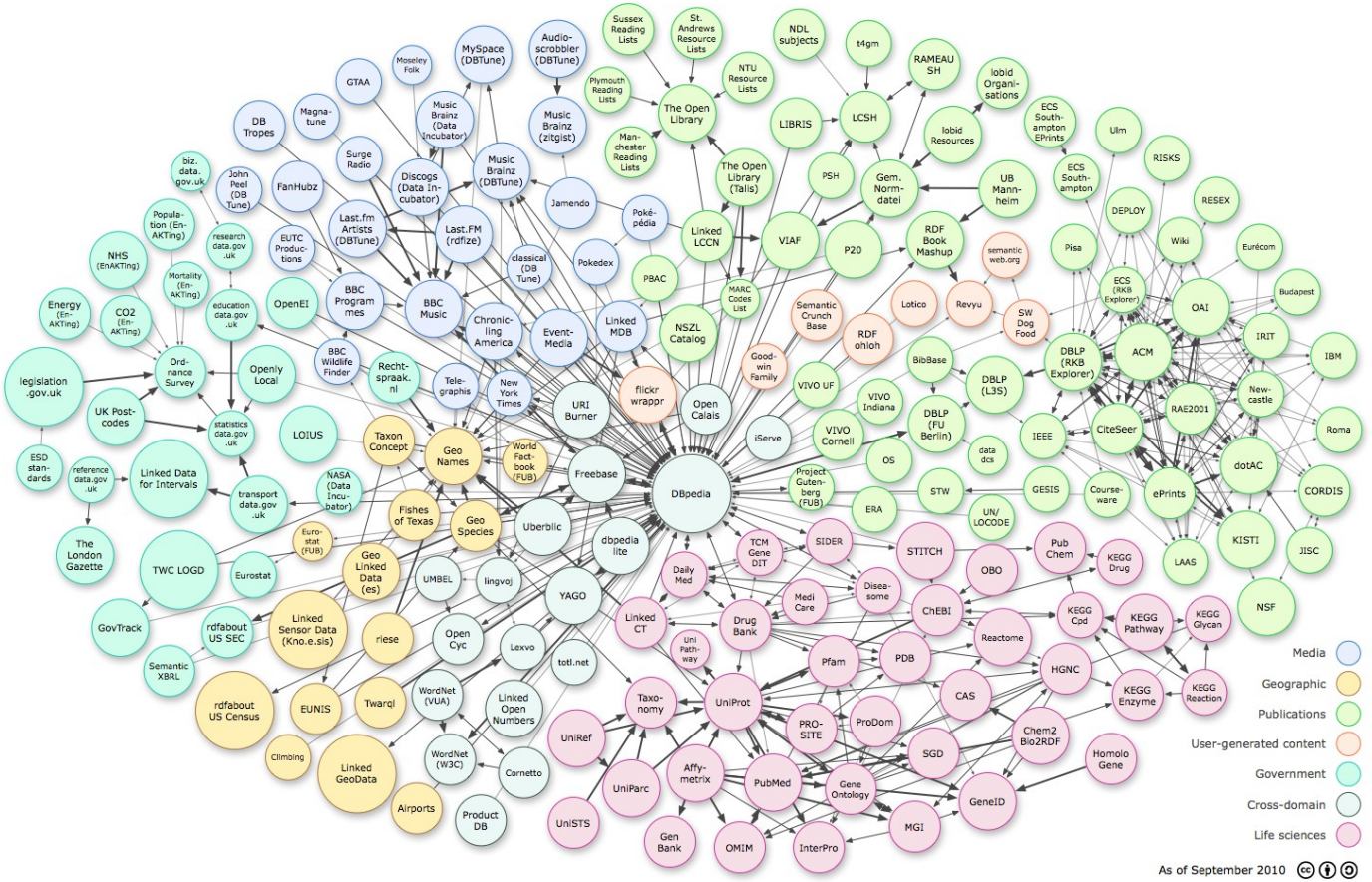
```
PREFIX …

ms:person a dbpedia:Physician.
ms:person address ms:addr.
ms:person name „Mark Smith“
ms:addr street „Main St. 14“
ms:addr city „Smalltown“
…
```

+



Semantic Web in a Nutshell

TECHNISCHE UNIVERSITÄT DARMSTADT



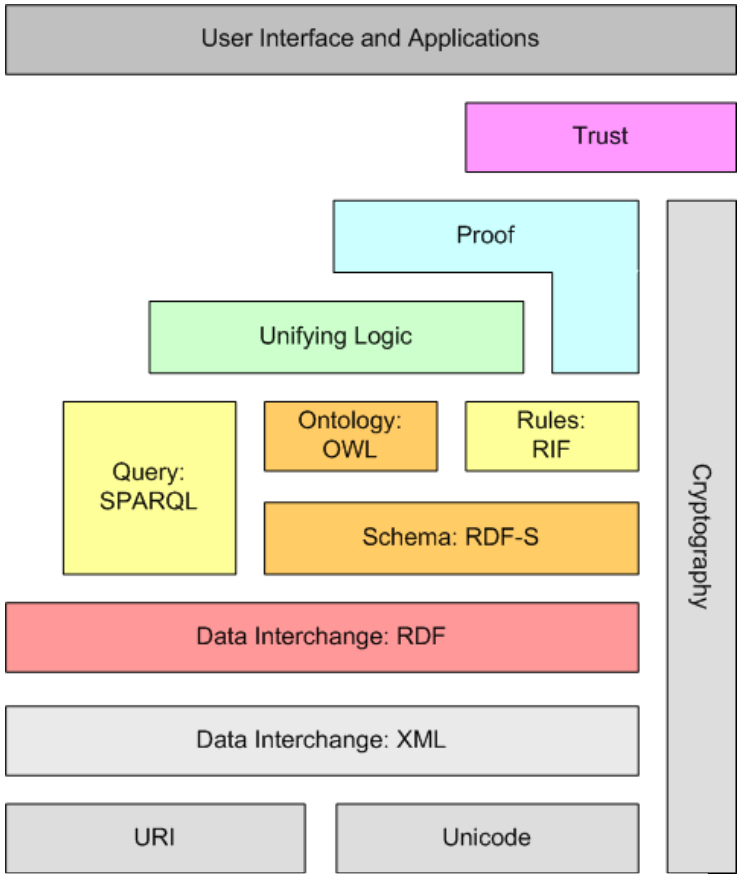
Linking Open Data cloud diagram,
by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

Semantic Web in a Nutshell

here be dragons...

Semantic-Web-Technologie

Technische Grundlagen



Berners-Lee (2009): *Semantic Web and Linked Data*
http://www.w3.org/2009/Talks/0120-campus-party-tbl/

Werbung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Vorlesung Semantic Web (3+1)
  - dieses Semester neu!
  - Dienstag 9:50 – 11:30
  - Donnerstag 8:55 – 10:35
  - in S202/C110

Semantic Web: Themen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Data Mining on Linked Open Data
  - Wie kann man Linked Open Data für "klassische" Data-Mining-Probleme nutzen?
- Drei Themen:
  - Statistical SPARQL
  - ProLOD und LiDDM
  - Automated Feature Generation from Linked Open Data

Semantic Web: Themen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ontology Learning
  - Aus Instanzdaten (A-box) automatisch auf die zugrunde liegende Ontologie (T-box) schließen
- Vier Themen:
  - Statistische Ansätze
  - ILP-basierte Ansätze (z.B. DL-Learner)
  - Regellern-basierte Ansätze (z.B. DL FOIL)
  - Information Extraction: Lernen von Ontologien aus Text

Semantic Web: Themen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Learning Queries
  - SPARQL-Abfragen sind für den Nutzer nicht leicht zu erstellen
  - Query by example: Nutzer gibt Beispiel-Ergebnisse, daraus wird die Abfrage gelernt
    - AutoSPARQL
  - Lernen von Zusammenhängen (Pathfinding)
    - Nutzer gibt Beispiele für einen Zusammenhang vor
    - die Abfrage, die den Zusammenhang erklärt, wird daraus gelernt

Semantic Web: Themen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ontology Matching
  - Verschiedene Ontologien werden im Semantic Web genutzt
  - Ontology Matching findet "Übersetzungen" zwischen Ontologien
  - Es gibt viele unterschiedliche Matching-Tools
    - Machine Learning: Trainieren eines Ensembles
    - Optimale Kombination von Ergebnissen

Literatur

- R. S. Michalski. On the Quasi-Minimal Solution of the Covering Problem. In (FCIP-69), pages 125–128, 1969.
- P. Clark and T. Niblett. The CN2 Induction Algorithm. Machine Learning, 3(4):261–283, 1989.
- W. W. Cohen. Fast Effective Rule Induction. In (ICML-95), pages 115–123, 1995.
- E. Frank and I. H. Witten. Generating Accurate Rule Sets Without Global Optimization. In (ICML-98), pages 144–151, 1998.
- J. R. Quinlan. Learning Logical Definitions from Relations. Machine Learning, 5:239–266, 1990.
- S. H. Muggleton. Inverse Entailment and Progol. New Generation Computing, 13(3,4):245–286, 1995.
- J. Fürnkranz. Separate-and-Conquer Rule Learning. Artificial Intelligence Review, 13(1):3–54, 1999.
- J. R. Quinlan and R. M. Cameron-Jones. Oversearching and Layered Search in Empirical Learning. In (IJCAI-95), pages 1019–1024, 1995.
- F. Janssen and J. Fürnkranz. A Re-evaluation of the Over-Searching Phenomenon in Inductive Rule Learning. In (SDM-09), pages 329–340, 2009.