Ontology-based data mining

Agnieszka Ławrynowicz

Institute of Computing Science
Poznan University of Technology

Kazimierz Dolny, April 27, 2010
I. Methods
- Introduction to data mining with ontologies
- Frequent pattern discovery
- Clustering of query results

II. Activities
- National projects
- International cooperation
  - Bari (Italy), Aberystwyth (UK)
  - EU FP7 project e-LICO (Data Mining Ontology, Multimedia Repositories Ontology)
- Dissemination: IRMLeS workshops
Background

- Volumes of digital data (Web application interfaces, bar code readers, bank transactions)
- Complex forms of data
  - Domain knowledge available (bioinformatics, medical informatics)

Knowledge discovery (data mining)

Knowledge representation (ontologies)

Directly handling complex data and (ontological) background knowledge in data mining process (Relational Data Mining, Inductive Logic Programming)
Knowledge discovery (KDD): non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data

- Data mining: application of automated techniques for pattern discovery and extraction

RDM/ILP:
- so far mostly Logic Programs, Datalog as representation language,
- recent trends in KR: OWL (description logics)
Frequent pattern discovery

- Given ontology $O$ and corresponding knowledge base $KB$
- Find a set of queries, $Q$ that capture the characteristics of $KB$ – frequent patterns

Example:
$q(x) :- \text{Client}(x), \text{Man}(x), \text{isOwnerOf}(x, y), \text{GoldCreditCard}(y)$
support: 20%

Józefowska J., Lawrynowicz A., Łukaszewski T., The role of semantics in mining frequent patterns from knowledge bases in description logics with rules, Accepted to *Theory and Practice of Logic Programming* journal, TPLP, Spring 2010
General idea: instead of aggregating the results based on values in tuples, aggregation is done based on their semantics (e.g. classes) expressed by ontologies.
The user Anna searches for natural monuments located in East-Central Europe. She would like to have the monuments grouped by their proximity.
The user Anna queries for a short weekend break offer. She would like to have the offers grouped by the type of destination (city, resort etc.). She would like also to have the information about the content of each group in terms of intensional descriptions.
Why Semantics of GROUP BY is not enough

a) Would be grouping by longitude and latitude what the user Anna actually expects?
   • GROUP BY semantics is to partition the results by identical values
   • no natural monuments share the same coordinates
   • one row for each value would be created ✗
Why Semantics of GROUP BY is not enough

a) Would be grouping by longitude and latitude what the user Anna actually expects?
   • GROUP BY semantics is to partition the results by identical values
   • no natural monuments share the same coordinates
   • one row for each value would be created

b) Consider names of towns as the instances of destination
   • GROUP BY would create one row for each name of a town (still too many groups)
Semantic aggregation

- CATEGORIZE BY ("semantic GROUP BY")
- CLUSTER BY (semantic clustering using data mining)

ASPARAGUS-- Automatic SPARQL query results AGgregation Using Semantics (system in early stage of development)

Ławrynowicz A., Grouping Results of Queries to Ontological Knowledge Bases by Conceptual Clustering, ICCCI 2009, Springer LNCS
Ławrynowicz A., Query Results Clustering by Extending SPARQL with CLUSTER BY, OTM Workshops 2009, Springer LNCS
National projects

• Frequent pattern discovery from knowledge bases in description logic with rules (2007-2009)
• Inductive reasoning on ontological knowledge bases (2009-2012)
International cooperation

Univ. of Aberystwyth – UK (prof. Ross D. King)
- Applications to systems biology
- Short visit, invited seminar, November 2007

Univ. of Bari - Italy (Claudia d'Amato, Nicola Fanizzi)
- Methods of data mining with ontologies, non-standard reasoning with description logics
- Short sabbatical period in Bari incl. series of invited seminars in November 2009

d'Amato C., Fanizzi N., Ławrynowicz A., Categorize By: Deductive Aggregation of Semantic Web Query Results, Accepted for publication in Proc. of Extended Semantic Web Conference'2010, Spinger, LNCS
EU FP7 e-LICO Objectives

- Infrastructure to support collaborative, data mining enabled experimental research
- Knowledge-driven, planner-based intelligent discovery assistant that will self-improve through meta-mining
Extensive use of **prior knowledge** in the mining process:

- **Data Mining:**
  - Ontology-based planning and workflow generation
  - Ontology-based meta-learning for algorithm and model selection

- **Domain** ontologies for pilot applications

Meta-learning over the full data mining process

= **meta-mining**
The e-LICO DM lab
e-LICO-Enlarged EU

- **October 2009:** ICT Call 5 closes with special objective for "new" EU countries (*Enlarged EU*) to join ongoing projects
- **May 2010:** Grant agreement signature date
- **1 June 2010:** Project extension starts
PUT ICS in e-LICO-Enlarged EU

• Extensive activities in Data Mining Ontology development (initial co-operation with Univ. of Geneva started December 2009)
• Collaboration on new pilot application domain - Digital Multimedia Repositories
  - Collaborative development of domain ontology
  - New tasks – recommender systems for videolectures.net
• Ontology based data mining methods:
  - frequent pattern discovery,
  - semantic kernels = “semantic” similarity measures (in meta-mining)
Data Mining Ontology and KB - overview

DMO: >700 concepts, ~100 properties
Data Mining Ontology – main concepts
Data Mining Ontology – open issues

- How to represent semantics of mathematical formulas?
  - Example: cost function
- part_of: single relation?, a hierarchy of specialized relations? re-using some ontology, e.g. RO – Relations Ontology?

- top ontologies
Digital Multimedia Repositories
Ontology - upcoming
IRMLeS workshops

- IRMLeS – Inductive Reasoning and Machine Learning on the Semantic Web
- co-located with major European Semantic Web conference (ESWC'2009, ESWC'2010)
- Organizing Committee: Claudia d'Amato, Nicola Fanizzi (Italy), Marko Grobelnik (Slovenia), Agnieszka Ławrynowicz (Poland), Vojtěch Svátek (Czech Republic)
- Proceedings published in CEUR.WS
- Post-workshop publication planned in new Semantic Web journal by IOS Press
Thank you!