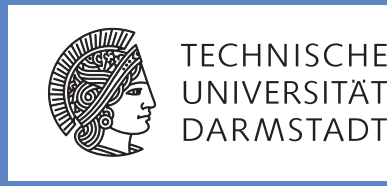
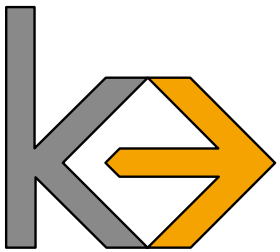
DBpedia Ontology Enrichment for Inconsistency Detection and Statistical Schema Induction



Presentation By Tung Do



1. Statistical Schema Induction
2. DBpedia Ontology Enrichment for Inconsistency Detection

Statistical Schema Induction: Introdution

TECHNISCHE UNIVERSITÄT DARMSTADT

- a statistical approach to the induction of expressive schemas from large RDF repositorie
- Discover hidden knowledge from ontological knowledge bases
- By Auer and Lehman ontologies derived from RDF repositories can also bring major benefits for theWeb of Data
- By providing conceptual descriptions of RDF graphs ontologies might facilitate, for instance, the discovery of links between disconnected data sets, or enable the detection of contradictory facts spread across the cloud of Linked Open Data.

Logical and statistical methods

TECHNISCHE UNIVERSITÄT DARMSTADT

- Logical method : as Inductive Logic Programming
  - Generation of highly axiomatized ontologies
- Statictical method : based on conceptual clustering
  - More scalable
  - Robust with repect to noisy or uncertain data

Organization

TECHNISCHE UNIVERSITÄT DARMSTADT

- Some overview of related work
- Introduce the EL profile of OWL 2
- Detail the implementation
- Evaluation on several real-world datasets

Inductive Logic Programming (ILP)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Derive logical theories from example and background knowledge
- ILP-based methods have successfully been applied to the problem of concept learning and ontology induction, e.g., by Cohen and Hirsh [10].
- Hellmann applied the DL-Learner to several RDF knowledge bases, in order to generate definitions of classes from the YAGO ontology
- Another particularly interesting approach has been proposed by Cimiano , who generate intentional descriptions of the factoid answers (e.g. sets of individuals) that are returned by queries to a given knowledge base.

Formal Concept Analysis (FCA) or Relation Exploration

TECHNISCHE UNIVERSITÄT DARMSTADT

- OntoComP developed by Baader supports knowledge engineers in the acquisition of axioms expressing subsumption between conjunctions of named classes
- A similar method for acquiring domainrange restrictions of object properties has been proposed later by Rudolph.

Association Rule

TECHNISCHE UNIVERSITÄT DARMSTADT

- Applied in the area of ontology matching as in the AROMA system
- Work by Parundekar , who consider containment relationships between sets of class instantiations for producing alignments between several linked data repositories, including DBpedia
- Determine the type of correspondence between a given pair of restriction classes by Parundekar rely on thresholds applied to measures of extensional overlap

OWL 2 EL

TECHNISCHE UNIVERSITÄT DARMSTADT

- Based on the description logic EL++ , reasoning services such as consistency and instance checking can be performed in time that is polynomial with respect to the number of axioms
- Description logics define concept descriptions inductively by a set of constructors , starting with a set N_c of concept (or class) names, a set N_r of role (or property) names, and a set N_I of individual names

OWL 2 EL

TECHNISCHE UNIVERSITÄT DARMSTADT

Name	Syntax	Semantics
Top	$\top$	Δ^x
Bottom	$\bot$	$\varnothing$
Conjunction	$C \bigsqcap D$	$C^x \cap D^x$
Existentical restrition	$\exists r.C$	$x \subset \Delta^x \vert \exists y \subset \Delta^x : (x,y) \in r^x$ $\wedge y \in C^x$
GCI	$C \sqsubseteq D$	$C^x \sqsubseteq D^x$
RI	$r_1 \circ \cdots \circ r_k \sqsubseteq r$	${r_1}^x \circ \cdots \circ {r_k}^x \sqsubseteq r^x$

Association rule Mining

TECHNISCHE UNIVERSITÄT DARMSTADT

- A very simple but useful form of implication patterns
- Framework was developed for large and sparse datasets such as transaction databases of international supermarket chains

$$supp(x) = \left|\{t_i \in D : X \subseteq t_i\}\right|$$

$$conf(A \Rightarrow B) = \frac{supp(A \cup B)}{supp(A)}$$

Association rule Mining

TECHNISCHE UNIVERSITÄT DARMSTADT

IRI	Comedian	Artist	Person	Airport	Building	Place	Animal
Jerry-Seinfeld	1	1	0	0	0	0	0
Black-Bird	0	0	0	0	0	0	1
Chris_Rock	1	1	1	0	0	0	0
Robin_Williams	1	0	1	0	0	0	0
JFK_Airport	0	0	0	1	1	1	0
Hancock_Tower	0	0	0	0	1	1	0
NewWark_Airport	0	0	0	1	1	1	0

Tabelle : Example of a transaction database in the context of the DBpedia dataset

SSI

TECHNISCHE UNIVERSITÄT DARMSTADT

- based on the assumption that the semantics of any RDF resource is revealed by patterns we can observe when considering the usage of this resource in the repository
- process of SSI :
  - Terminology (collection and create a set S of realation database)
  - Association Rule Mining (create transaction table form S and use it to generate Association Rule)
  - Ontology Construction (based on new Rule for build Ontology)

SSI

TECHNISCHE UNIVERSITÄT DARMSTADT

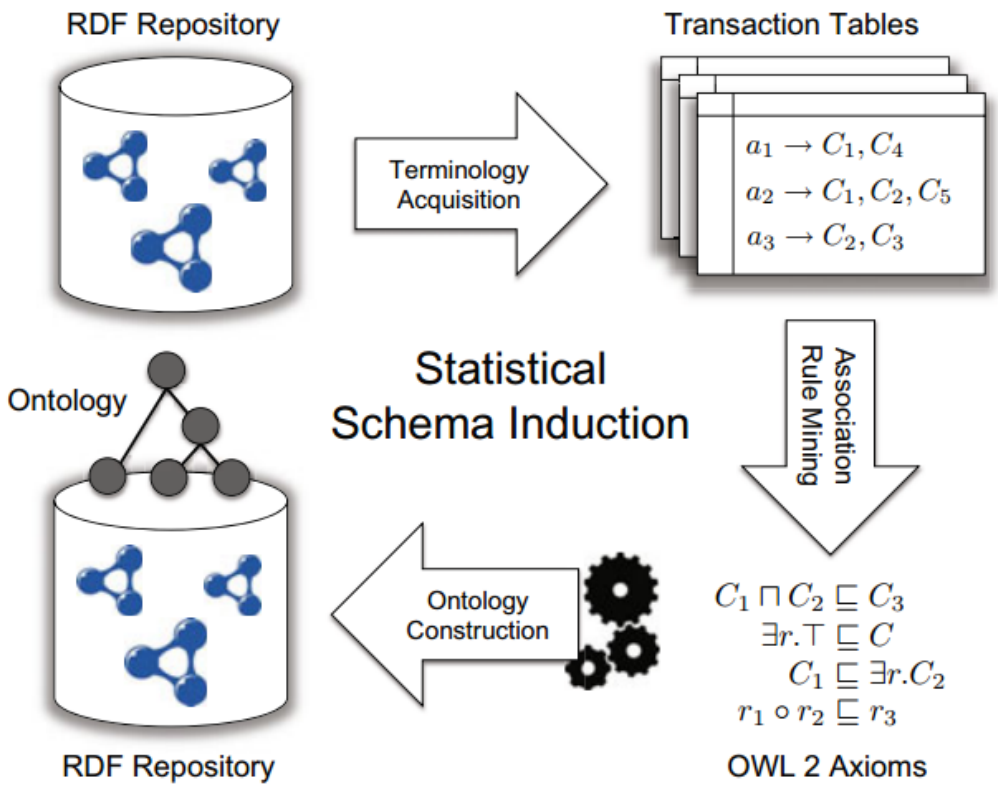


Abbildung : Worfklow of the Statistical Schema Induction framework

Terminology Acquisition

TECHNISCHE UNIVERSITÄT DARMSTADT

- Name classes:
  - Gather information about those resouce which are likely to represnt classes C
    - every object of an rdf:type statement is a class
  - Consider the use of sample heristics
- Object properties : collect the names of all those RDF resources which we assume to represent object properties *r*
  - Every predicate of an RDF triple which belongs to the DBpedia namespace and whose object is linked to another resource by means of an rdf:type statement is considered an object property
- Class expression : turn to comple class and property expressions
- Property chains : acquire transitivity axioms for all the predicates

Association Rule Mining

TECHNISCHE UNIVERSITÄT DARMSTADT

$C \sqsubseteq D$	$a \rightarrow C_1, ..., C_n$ for $a \in N_1$	$\{C_i\} \Rightarrow \{C_j\}$
$C \sqcap D \sqsubseteq E$	$a \rightarrow C_1, ..., C_n$ for $a \in N_1$	$\{C_i, C_j\} \Rightarrow \{C_k\}$
$D \sqsubseteq \exists r.C$	$a \rightarrow C_1, ..., C_l, \exists r_1.C_1 1, ..., \exists r_m.C_{mn}$ for $a \in N_1$	$\{C_k\} \Rightarrow \{\exists r_j.C_j k\}$
$\exists r.C \sqsubseteq D$	$a \rightarrow C_1, ..., C_l, \exists r_1.C_1 1, ..., \exists r_m.C_{mn}$ for $a \in N_1$	$\{\exists r_j.C_j k\} \Rightarrow \{C_i\}$
$\exists r.\top \sqsubseteq C$	$a \rightarrow C_1, ..., C_l, \exists r_1.\top, ..., \exists r_m.\top$ for $a \in N_1$	$\{\exists r_j.\top\} \Rightarrow \{C_i\}$
$\exists r^{-1}.\top \sqsubseteq C$	$a \rightarrow C_1, ..., C_l, \exists {r_1}^{-1}.\top, ..., \exists {r_m}^{-1}.\top$ for $a \in N_1$	$\{\exists {r_j}^{-1}.\top\} \Rightarrow \{C_i\}$
$r \sqsubseteq s$	$(a, b) \rightarrow r_1, ..., r_n$ for $(a, b) \in N_1 x N_1$	$\{r_i\} \Rightarrow \{r_j\}$
$r \circ r \sqsubseteq r$	$(a, b) \rightarrow r_1, ..., r_n, r_1 \circ r_1, ..., r_n \circ r_n$ for $(a, b) \in N_1 x N_1$	$\{r_i \circ r_i\} \Rightarrow \{r_j\}$

Ontology Constrution

TECHNISCHE UNIVERSITÄT DARMSTADT

- An initially empty or to an existing OWL ontology that we would like to refine
- Sort all of the generated axioms in descending order based on their certainty values
- Add them to the ontology one by one, checking the coherence of the ontology after the addition of each axiom.

Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Both of these experiments were carried out on a an AMD 64bit DualCore computer with 2,792 MHz and 8 GB RAM. The Java-based implementation of our approach makes use of various publicly available libraries for database access (MySQL 5.0.51), ontology management (Pellet 2.2.1 and OWL API 3.0.0) and Linked Data querying (Jena 2.6.3).

1110	1325	6293	0	1	144			
4056	4665	1146	1146	6330	6973	64	185	
1146	6330	6973	64	68	141			
3235	6668	6769	3242	5049	6673	3907	2	66
1110	1325	9293	0	73	144			

Tabelle : Textual serialization of a transaction table

Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

(a) Without support threshold

τ	# axioms	recall	precision	F_1 score
1.0	365	0.997	0.992	0.995
0.9	373	0.997	0.971	0.983
0.8	381	0.997	0.950	0.973

(b) With support threshold of 10

τ	# axioms	recall	precision	F_1 score
1.0	339	0.926	0.991	0.957
0.9	347	0.926	0.968	0.947
0.8	354	0.926	0.949	0.937

Abbildung : Recall, precision and F_1 values for **subsumption axioms** between atomic classes for varying thresholds on the confidence values

(a) Without support threshold

τ	# axioms	recall	precision	F_1 score
1.0	950	0.900	0.808	0.852
0.9	1143	0.946	0.655	0.774
0.8	1181	0.946	0.683	0.793

(b) With support threshold of 10

τ	# axioms	recall	precision	F_1 score
1.0	821	0.821	0.821	0.805
0.9	1036	0.576	0.576	0.682
0.8	1092	0.558	0.558	0.670

Abbildung : Recall, precision and F_1 values for **domain restriction axioms** between atomic classes for varying thresholds on the confidence values

DBpedia

TECHNISCHE UNIVERSITÄT DARMSTADT

- The DBpedia ontology was created by a manual mapping of 1,055 Wikipedia infobox templates to 259 named classes. Besides these classes, the ontology comprises 602 object properties, 674 datatype properties, 257 explicit subsumption axioms as well as 459 domain and 482 range restrictions. 1,477,796 of the roughly 3.4 million ẗhings̈(i.e. RDF resources representing Wikipedia articles) are explicitly classified with regard to the DBpedia ontology.
- The time needed to compute the association rules was less than 5 seconds for the largest transaction table, which confirms the scalability of the Apriori algorithm to the large Linked Data repositories

DBpedia

TECHNISCHE UNIVERSITÄT DARMSTADT

(a) Without support threshold

τ	# axioms	recall	precision	F_1 score
1.0	401	0.392	0.666	0.494
0.9	740	0.712	0.655	0.682
0.8	868	0.790	0.620	0.695

(b) With support threshold of 10

τ	# axioms	recall	precision	F_1 score
1.0	206	0.258	0.854	0.396
0.9	740	0.580	0.512	0.544
0.8	840	0.658	0.604	0.630

Abbildung : Recall, precision and F_1 values for range restriction axioms for varying thresholds on the confidence values

Other Datasets

TECHNISCHE UNIVERSITÄT DARMSTADT

- Without any changes to our implementation, we were able to compute appropriate transaction tables for five subsets of data.gov.uk each of these subsets corresponding to a public sector:8 reference, eduction, ordnance, transport and finance.
- Then, we use these restrictions in order to c̈lassifyäll of the resources in the RDF graph . Without applying this strategy, we obtained 64 classes and 47 axioms for education , 62 classes and 137 axioms for transport , 20 classes and 17 axioms for reference , 29 classes and 3 axioms for ordnance , as well as 41 classes and 0 axioms for finance , where äxiomsr̈efers to explicit subsumption axioms ($C \sqsubseteq D$).

Conclusion

TECHNISCHE UNIVERSITÄT DARMSTADT

- statistical schema induction as a means to generating ontologies from RDF data
- As future work we envision, for example, an adaptation of our approach to more expressive description logics.
- Furthermore, we would like to facilitate a more efficient construction of the transaction tables by appropriate sampling strategies or a Map-Reduce framework for distributed computation.
- changes to be strictly monotonic, the necessary adaptations to the transaction tables will be linear in time, and efficient algorithms for mining association rules could suggest appropriate ontology refinements within a few seconds at most.

DBpedia Ontology Enrichment for Inconsistency Detection

TECHNISCHE UNIVERSITÄT DARMSTADT

- the provided data is only valuable if it is accurate and without contradictions.
- enable the detection of inconsistencies
- DBpedia data often is error-prone that inconsistencies are detected during the extraction.
- the automatic extraction based on Wikipedia resources that have been created by a large number of non-expert users its data is partly incorrect or incomplete.
- By applying the enriched ontology during the extraction process it is possible to induce contradictions that point to incorrect facts

Organization

TECHNISCHE UNIVERSITÄT DARMSTADT

- Recapitulates previous work in the field of error detection and correction in Linked Data as well as ontology enrichment
- The proposed methods applied for enriching the DBpedia ontology
- the detection of inconsistencies, which has been integrated into the DBpedia Extraction Framework

Related work

TECHNISCHE UNIVERSITÄT DARMSTADT

- Hogan have focused on different types of errors referring to accessibility, syntactical correctness, and consistency of published RDF data
- the detection of inconsistencies within DBpedia is highlighted, which is achieved by automatic semantic enrichment of the underlying ontology
- ORE is using that framework as a foundation and learns axioms, which express a subclass relationship or a class equivalence. (an integrated reasoner)
- with annotations, which comprise a declaration about the correctness and the relevance of the axioms .
- In four different approaches for handling inconsistencies in changing ontologies have been surveyed:
  - the evolution of a consistent ontology .
  - the reparation of inconsistencies .
  - reasoning in the presence of inconsistencies .
  - multi-version reasoning .
- Crowdsourcing : apply human intelligence for the detection of errors in knowledge bases .

Level of erors

TECHNISCHE UNIVERSITÄT DARMSTADT

- syntactic errors : can be detected with the help of a simple RDF parser/validator
- logical errors : can be identified with the help of a reasoner
- semantic errors comprise facts that are not corresponding to facts in the real world.

Example

TECHNISCHE UNIVERSITÄT DARMSTADT

dbp:2666_%28novel%29 dbo:publisher dbp:Barcelona

- to transform semantic errors into logical ones by extending the axioms of the underlying ontology
  dbp:2666_%28novel%29 dbo:publisher dbp:Barcelona .
  dbo:publisher rdfs:range dbo:Company .
  dbp:Barcelona rdf:type dbo:Settlement .
- By adding the disjointness axiom
  dbo:Company owl:disjointWith dbo:Settlement .

Abbildung : Extract of the infobox in the article 2666

Property domain restrictions

TECHNISCHE UNIVERSITÄT DARMSTADT

- domain restrictions have to be specified explicitly.

$$md_{p,c} = \frac{\left|\{(s\ p\ o) : (s\ p\ o) \in KB \wedge (s\ a\ c) \in KB\}\right|}{\left|\{(s\ p\ o) : (s\ p\ o) \in KB \wedge (s\ a\ d) \in KB \wedge d \neq T\right|}$$

- $md_{p,c}$: indicates whether class c is the domain of property p
- $KB = \{(s\ p\ o) : s \in E \cup C \circ p \in P \circ o \in E \cup L \cup C\}$
- $(s\ p\ o) \in KB$: subject $s \in E$ belongs to the class c, relatively to the number of triples $(s\ p\ o) \in KB$, whose subject $s \in E$ belongs to any class more specific than owl:Thing (abbreviated T in the equations
- The equations use the abbreviation $(s\ a\ c)$ for expressing the fact that the entity $s \in E$ belongs to the class c.
- The appropriate threshold $T_{md} = 0.96$ has been determined via a randomized analysis

Property rangge restricticons

TECHNISCHE UNIVERSITÄT DARMSTADT

$$mr_{p,c} = \frac{\left|\{(s\ p\ o) : (s\ p\ o) \in KB \wedge (o\ a\ c) \in KB\}\right|}{\left|\{(s\ p\ o) : (s\ p\ o) \in KB \wedge (o\ a\ d) \in KB \wedge d \neq T\right|}$$

- $mr_{p,c}$: indicates whether c is the range of an object property p
- The threshold T_{mr}= 0.77, which has been investigated by means of a randomized analysis, seems appropriate.

Class disjointness axiom

TECHNISCHE UNIVERSITÄT DARMSTADT

- documents contain equivalent terms , entities of similar classes occur more frequently with the same properties
- the weight of a term in a document for a given set of documents can be determined based on the term frequencyinverse document frequency (TF-IDF).

$$w_{c,p} = pf_{c,p} * icf_p$$

- the absolute frequency of a property p along with the entities of the class c

$$pf_{c,p} = \left|\{(s\ p\ o) : (s\ p\ o) \in KB \wedge (s\ a\ c) \in KB\}\right|$$

$$wcf_{c,p} = \begin{cases} 1 + \log pf_{c,p} & \text{if } pf_{c,p} > 0 \\ 0 & 0 \end{cases}$$

Class disjointness axiom

TECHNISCHE UNIVERSITÄT DARMSTADT

- General relevance of a property p for the complete knowledge base

$$icf_p = \log \frac{|C|}{|\{c : c \in C \wedge (s\ p\ o) \in KB \wedge (s\ a\ c) \in KB\}|}$$

- The similarity value $sim_{ci,cj}$ of two classes c_i and c_j is normalized between 0 und 1, since $w_{c_i,p_k} \geq 0$ and $w_{c_j,p_k} \geq 0$ holds for all p_k.

$$sim_{c_i,c_j} = \frac{\overrightarrow{v_{c_i}} * \overrightarrow{v_{c_j}}}{|\overrightarrow{v_{c_i}}||\overrightarrow{v_{c_j}}|} = \frac{\sum_{k=1}^{n} w_{c_i,p_k} w_{c_j,p_k}}{\sqrt{\sum_{k=1}^{n} w^2{}_{c_i,p_k}} \sqrt{\sum_{k=1}^{n} w^2{}_{c_j,p_k}}}$$

- A random sample has pointed out that $\tau_{sim} = 0.17$ seems to be an appropriate threshold.

Detection of inconsistences

TECHNISCHE UNIVERSITÄT DARMSTADT

- checking the consistency of the range of a property in an RDF triple needs the information about the rdf:type of the object
- Solution variants
  - (D1) Map the template property onto another ontology property.
    - the inconsistency results from an incorrectly used property
  - (D2) Remove the classes' disjointness axiom.
    - the class that represents the rdf:type of the subject and the class
  - (D3) Change the domain of the ontology property to owl:Thing
    - the inconsistencies might be caused by the fact that the ontology property is generic
- Violation of range of property
  - (R1) Create a link to the appropriate article in th articles infobox.
    - a range-violation is caused by the fact that a linked article in a value of an infobox is confused with the actual article
  - (R2) Delete the value or associate the value with another template property
    - information does not fit to the template property within an infobox
  - (R3) Map the template property onto another ontology property.
  - (R4) Remove the classes' disjointness axiom.
  - (R5) Change the range of the ontology property to owl:Thing.

Enrichment

TECHNISCHE UNIVERSITÄT DARMSTADT

- For all properties occurring in the ABox a domain restriction has been determined with the aid of the metric $md_{p,c}$ - either a class of the DBpedia ontology or the class owl:Thing. Out of the 1,363 properties 5% are randomly chosen und the correctness of their classification is checked

N	n	t_p	f_p	$\widehat{pr}$	95% confidence interval
1,363	68	67	1	0.985	[0.957, 0.999]

- In Section 3 a range restriction has been assigned to 592 properties. Approximately 10% of these properties have been randomly chosen for manually verification of the classification.

N	n	t_p	f_p	$\widehat{pr}$	95% confidence interval
592	59	51	8	0.864	[0.781, 0.948]

Enrichment

TECHNISCHE UNIVERSITÄT DARMSTADT

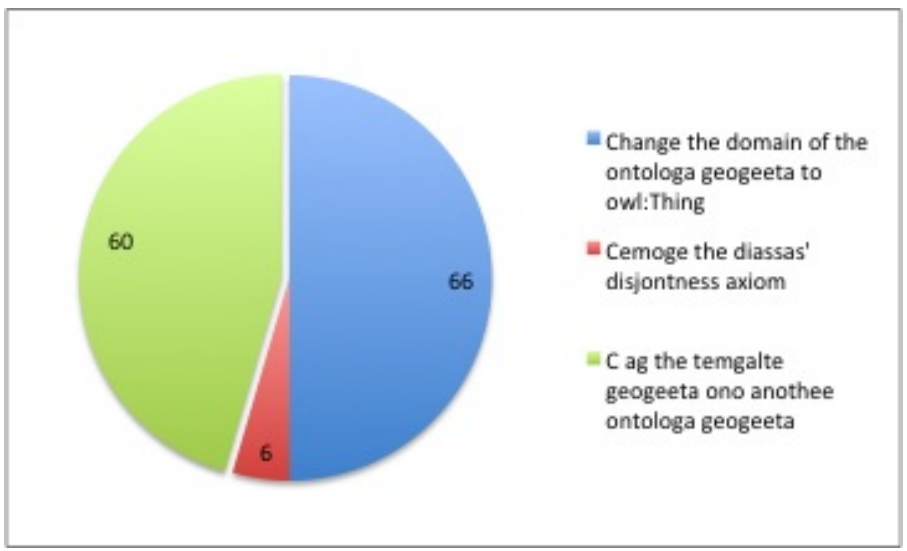
- Out of the 37,091 class pairs that have been declared as disjoint classes approximately 0,5% of the pairs are taken as a random sample. Subsequently their classification is checked.

N	n	t_p	f_p	$\widehat{pr}$	95% confidence interval
37091	185	183	2	0.989	[0.974, 1.0]

Detection

TECHNISCHE UNIVERSITÄT DARMSTADT

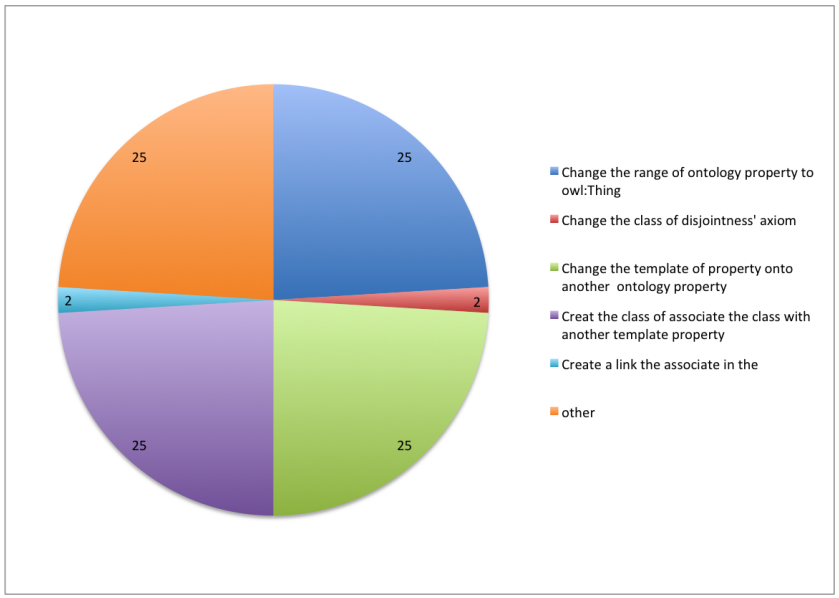
- The consistency check examined 3,110,392 entities, whereas the majority of 3,060,898 resources showed to be consistent. The remaining 49,494 entities have been classified as inconsistent having 60,602 inconsistencies. In 12,218 cases the inconsistency results from a domain restriction violation of a used property, 40,404 inconsistencies result from range restriction violations.



Detection

TECHNISCHE UNIVERSITÄT DARMSTADT

- the ratio of the different suggestions, which have been applied to eliminate domain restriction violations.



- the ratio of the different suggestions, which are applied in order that range restriction violations are eliminated.

Conclusion

TECHNISCHE UNIVERSITÄT DARMSTADT

- Identify inconsistent triples during the extraction process of the DBpedia dataset which may lead to a higher quality extraction
- The applied methods performed with reasonably high precision, which allows to use the enriched ontology
- Minimizes the number of remaining inconsistencies, to support the decision of the user.