UNIVERSITÄT
KOBLENZ · LANDAU
Institut WeST

FB 4
Informatik

Explicit and Implicit Schema Information on the Linked Open Data Cloud: Joined Forces or Antagonists?

Ansgar Scherp
Thomas Gottron
Malte Knauf
Stefan Scheglmann

Nr. 6/2012

Arbeitsberichte aus dem Fachbereich Informatik

Universität Koblenz-Landau • Campus Koblenz • Universitätsstraße 1 • 56070 Koblenz
Telefon +49 261 287-0 • http://www.uni-koblenz.de

Arbeitsberichte des Fachbereichs Informatik

ISSN (Print): 1864-0346
ISSN (Online): 1864-0850

Kontaktdaten der Verfasser

Ansgar Scherp, Thomas Gottron, Malte Knauf, Stefan Scheglmann
Institut WeST
Fachbereich Informatik
Universität Koblenz-Landau
Universitätsstraße 1
D-56070 Koblenz
E-Mail: scherp@uni-koblenz.de, gottron@uni-koblenz.de, mknauf@uni-koblenz.de, schegi@uni-koblenz.de

Explicit and Implicit Schema Information on the Linked Open Data Cloud: Joined Forces or Antagonists?

Thomas Gottron, Malte Knauf, Stefan Scheglmann, Ansgar Scherp

WeST – Institute for Web Science and Technologies
University of Koblenz-Landau
56070 Koblenz, Germany
{gottron,mknauf,schegi,scherp}@uni-koblenz.de

Abstract Schema information about resources in the Linked Open Data (LOD) cloud can be provided in a twofold way: it can be explicitly defined by attaching RDF types to the resources. Or it is provided implicitly via the definition of the resources' properties. In this paper, we analyze the correlation between the two sources of schema information. To this end, we have extracted schema information regarding the types and properties defined in two datasets of different size. One dataset is a LOD crawl from TimBL's FOAF profile (11 Mio. triple) and the second is an extract from the Billion Triples Challenge 2011 dataset (500 Mio. triple). We have conducted an in depth analysis and have computed various entropy measures as well as the mutual information encoded in this two manifestations of schema information. Our analysis provides insights into the information encoded in the different schema characteristics. It shows that a schema based on either types or properties alone will capture only about 75% of the information contained in the data. From these observations, we derive conclusions about the design of future schemas for LOD.

1 Introduction

Schematic information about semantic data on the Linked Open Data (LOD) cloud is given in a twofold way: explicitly by providing the type of a resource and implicitly via the definition of its properties. These two manifestation of schematic information are to a certain extent redundant, i.e., certain resource types entail typical properties and certain properties occur mainly in the context of particular types. For instance, we would expect a resource of type foaf:Person to have the properties foaf:name or foaf:age. Likewise, we can assume a resource with the property skos:prefLabel to be of type skos:Concept.

Schematic information over LOD is used for various purposes, such as indexing [8], data analytics [9], query optimization [10] or type prediction. Thus, it is an important question to which degree explicit and implicit schema information is correlated, i.e., to which extend the use of RDF types and properties appear together to describe resources. A high correlation of explicit and implicit schema information corresponds to redundant information – a fact which can be exploited, for instance, when indexing the LOD cloud and providing a central lookup table for LOD sources. One application in this context is the opportunity to compress a schema based index for LOD as motivated and requested by Neumann and Weikum [11]. More even, it is of interest, which schematic information actually needs to be extracted from the Linked Open Data cloud and which information might be inferred[1]. This leads us to the overall question to which extent the explicit schema information provided by RDF types coincides with the implicit schema information of the properties used in the LOD cloud. A fundamental prerequisite to answer this question is the availability of a reliable schema extracted from the LOD cloud that takes into account both explicit and implicit schema information. With our SchemEX approach [8,7], we can compute such a schema for huge amounts of RDF triples in an efficient manner. Thus, the SchemEX schema can be used to investigate the degree of correlation between RDF types and properties.

As the discussion of the related work in the subsequent section shows, such an investigation as presented in this paper has—to the best of our knowledge—not been done before. We will briefly introduce our SchemEX approach in Section 3, as it forms the basis for deriving a schema from LOD cloud data. In Section 4, we introduce a probabilistic schema distribution model. Based on this model, we identify different metrics that will be computed on the SchemEX schema. The metrics comprise different types of entropy as well as the mutual information. In Section 5, we describe the datasets for our analysis as well as the computation of the metrics. The results of our investigation are shown in Section 6. We discuss the results in Section 7 and draw some conclusions regarding implications for future LOD schema. In summary, we can say that about 75% of the properties and the types used in the LOD cloud are correlated. Thus, a schema for LOD should not be build on either explicit or implicit schema information only and should ideally be composed of a mixture of both. In addition, we can say that there are several highly related sets of RDF types and sets of properties, allowing a prediction from one to the other and vice versa.

2 Related Work

One application where schematic information can be of value is query optimization. Neumann and Moerkotte [10] employ so-called *characteristic sets*, which basically classify RDF resources by the correlation of their (outgoing) predicate links. Knowledge about these sets allows for quite precise estimates of the result cardinality of join operations. Further insights into the correlation between properties in an RDF graph were not necessary. Neither were explicit schema information provided in form of RDF types considered. A similar approach is presented by Maduko et al. [9]. Here the focus was on efficient approaches to estimate subgraph frequencies in a collection of connected graphs or a graph database. Also these subgraph frequencies could be used to determine the cardinality of intermediate join results during query optimizations. In [4] Harth et al. propose an approximative approach to optimize queries over multiple distributed LOD sources. They estimate an index structure (a QTree) over the sources, which is used to determine the contribution to the query results of the single sources.

Several tools aim at providing statistics for the LOD cloud. LODStats [2] is a tool and framework for computing 32 different statistics on Linked Open Data such as those

[1] Inference here can be realized in both ways: semantically or statistically.

covered by the Vocabulary of Interlinked Datasets (VoID) [1]. The tool provides descriptive statistics such as the frequencies of property usage and datatype usages, the average length of literals, or counting the number of namespaces appearing at the subject URI position [2]. LODStats operates on a single triple pattern, i.e., it does not provide statistics of, e.g., star patterns or other (arbitrary) graph patterns. However, it covers more complex schema-level characteristics like the RDFS subclass hierarchy depth [2]. Overall, analysis of the correlating use of different properties, RDF types, or the common appearance of properties and types like we investigate is out of scope. Also make-void[2] computes VoID-statistics for a given RDF file. These statistics usually contain information about the total number of triples, classes, properties, instances for each class, the uses of each property and the number of triples that link a subject on one domain to an object on another domain. Another framework for statistic generation on RDF data is RDFStats[3]. In contrast to make-void, RDFStats can also operate on SPARQL endpoints and uses a different vocabulary for its statistics.

Hogan et al. have conducted an empirical study to investigate the conformance of linked data sources with 14 different linked data principles [6]. Among others, the authors analyzed how different class terms and property terms of vocabularies are re-used and mixed by the linked data providers. As metric, the authors apply the number of unique namespaces used by the respective data providers and provide a ranked list in terms of top-5 and bottom-5 data providers. Detailed investigations about the correlation of class terms and property terms of different (or the same) vocabularies are out of scope of their work. In a similar fashion, Bizer et al. have analyzed the LOD cloud for its compliance with the Linked Data principle using nine different criteria. They provide statistics such as the LOD cloud diagram or the inter-linkage of the datasets in the LOD cloud.[4]

3 The SchemEX Index as Basis for the Analysis

The purpose of SchemEX [7,8] is to link schematic information to data sources which provide resources conforming to this schema. Data sources are, e.g., static RDF documents and SPARQL endpoints [5]. The central concepts of SchemEX are Typeclusters (TC) and Equivalence classes (EQC). A TC contains all data sources which provide resources conforming to a well defined set of types/classes. The EQC divide the data sources in each TC into disjoint subsets, defined by the set of properties the instances have and in which TC the object of the triple lies. An overview of the information contained in a SchemEX index is shown in Figure 1.

It is important to notice that data sources can occur in several TC or EQC as they typically describe more than one and – in particular – different kinds of resources. However, different occurrences of a data source conform to different (in particular disjoint) sets of resources.

Noteworthy about SchemEX is, that it can be computed very efficiently and for large datasets using a stream-based approach. In this case, the analytical component is operat-

[2] https://github.com/cygri/make-void last visit 22 June 2012
[3] http://rdfstats.sourceforge.net/ last visit 22 June 2012
[4] http://www4.wiwiss.fu-berlin.de/lodcloud/state/

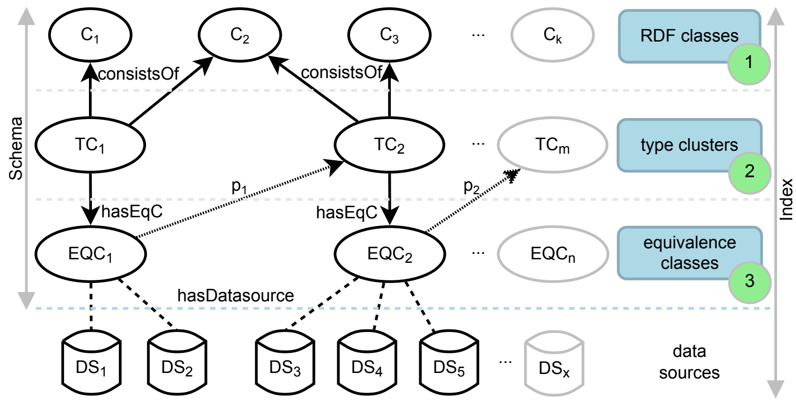


Figure 1. SchemEX index structure with three layers leveraging RDF typings and property sets

ing in a single pass fashion over a set of RDF triples. By using a windowing technique, it is possible to obtain a very accurate schema of the processed data using commodity hardware. However, the windowing technique entails a certain loss of schema information. The extent of this loss has been analyzed in detail in [7]. We will however look at the impact the loss has on the analysis performed in this paper by computing and comparing our metrics on a smaller dataset once for a lossless generated schema and once on a schema produced in the stream-based approach.

4 Probabilistic Schema Model and Metrics

Schema information on the LOD cloud can be provided explicitly by the use of RDF type properties. There are no (practical) boundaries to the number of types that can be attached to a resource. In practice, we can observe resources which have no type as well as resources with several hundred types. In addition, schema information can be provided implicitly by the properties used to describe a resource. These properties connect one resource to another resource or a literal value. In this way, they describe the type of a resource by its relations. Again, it is possible to observe resources which have no relation (beyond a type description) as well as resources with hundreds of properties.

The goal of the analysis in this paper is to measure and quantify the correlation between explicit schema information given by RDF types and implicit schema information provided by the used properties. To this end, in Section 4.1 we first introduce a probabilistic model for the occurrence of types and properties of resources. This allows us to measure the information contained in a schema based on types, properties or both. In order to do so, we apply different metrics such as entropy of marginal distributions, conditional entropy and mutual information in Section 4.2. Finally, we estimate the probabilities for the appearance of specific types and properties for the case of our SchemEX index in Section 4.3. These estimations are used as basis for our computations and analyses of linked open data in Section 5.

4.1 A Probabilistic Distribution for Types and Properties

We are interested in two observations about the resources on the LOD cloud: their types and their properties. To be more specific, we are interested in combinations of types and combinations of properties. A particular combination of types is a set of types attached to a resource. The space of all possible combinations therefore is the power set $\mathcal{P}(\mathit{Classes})$ of all class types in the data. While the power set itself is a huge set, we can actually restrict ourself to the subset $TS \subset \mathcal{P}(\mathit{Classes})$ of actually observed combinations of RDF types in the LOD cloud. For a given resource, we can now observe $t \in TS$ which corresponds to a set of types (e.g., the set $\{$foaf:Person, dbpedia:Politician$\}$).

Likewise, the properties observed for a resource is a combination of all possible properties. Accordingly here we deal with an element from the power set $\mathcal{P}(\mathit{Properties})$ of all observed properties. Again, we only need to consider the subset PS of actually occurred property sets. For an individual resource, we observe $r \in PS$ which corresponds to a set of its properties[5] (e.g., the set $\{$foaf:familyName, foaf:givenName, dbpedia:spouse$\}$).

To model the joint distribution of type sets and property sets, we introduce two random variables T and R. These take as values the elements in TS and PS, respectively. Both random variables are of discrete nature and their joint distribution can be characterized by:

$$P(T = t, R = r) = p(t, r) \tag{1}$$

Where $p(t, r)$ is the probability for a randomly chosen resource to observe the concrete set t of attached types and the set r of properties. Based on this joint distribution, we can also identify the marginal distributions of T and R:

$$P(T = t) = \sum_{r \in R} p(t, r) \quad , \quad P(R = r) = \sum_{t \in T} p(t, r) \tag{2}$$

4.2 Metrics of Interest

For analyzing the LOD cloud, we are interested in several characteristics of the joint distribution $P(T, R)$ introduced above. The main questions that we want to answer are:

(a) How much information is encoded in the type set or property set of a resource on a global scale?
(b) How much information is still contained in the properties, once we know the types of a resource?
(c) How much information is still contained in the types, once we know the properties of a resource?
(d) To which degree can one information (either properties or types) explain the respective other?

[5] Please note, we use the letter r for sets of properties (inspired by the term relation), as p will be used to denote probabilities.

To answer these questions, we introduce appropriate metrics that can be applied to compute the joint distribution of type sets and property sets. All our metrics are based on the *entropy* of probabilistic distributions [12], the standard concept to measure information.

Entropy of the Marginal Distributions: To answer the question of (a) how much information is encoded in the type or property set of a resource, we need to look at the marginal distributions. These provide us with the probability of a certain resource to show a particular set of types or properties. The entropy of the marginal distributions of T and R is defined as:

$$H(T) = -\sum_{t \in T} P(T = t) \cdot \log_2 \left(P(T = t) \right) \tag{3}$$

$$H(R) = -\sum_{r \in R} P(R = r) \cdot \log_2 \left(P(R = r) \right) \tag{4}$$

The values $H(T)$ and $H(R)$ give us an idea of how much information is encoded in the sets of types or properties of the resources. A higher value corresponds to more information, which in turn means that the sets of types or sets of properties appear more equally likely. To be more concrete: a value of 0 indicates that there is no information contained. For instance, a value of $H(T) = 0$ would indicate that all resources have exactly the same set of types (likewise for $H(R) = 0$). A maximal value, instead, is reached when the distribution is an equal distribution, i.e., each set of types or properties is equally likely. This fact also allows for normalizing the entropy values by:

$$H_0(T) = \frac{H(T)}{H^T_{\max}} = \frac{H(T)}{\log_2(|T|)} \quad , \quad H_0(R) = \frac{H(R)}{H^R_{\max}} = \frac{H(R)}{\log_2(|R|)} \tag{5}$$

The normalized entropy value ranges between 0 and 1 and indicates whether the distribution is closer to a degenerated or a uniform distribution.

Conditional Entropy: The question (b), how much information is still contained in the properties, once we know the types of a resource implies a conditional probability and, thus, a conditional entropy. We have to take a look at the distribution of the property sets given that we already know the types of a resource. The entropy in this case (i.e., the conditional entropy) conveys how much information is still in the additional observation of the properties. Again, if the set of types perfectly defines the set of properties to expect, there would be no more information to be gained. Thus, the conditional entropy would be zero. If, instead, the type was virtually independent from the properties, we would expect the marginal distribution of the properties and its according entropy. Formally the conditional entropy for a given type set t is given as:

$$H(R|T = t) = -\sum_{r \in R} P(R = r|T = t) \log_2 \left(P(R = r|T = t) \right) \tag{6}$$

$$= -\sum_{r \in R} \frac{p(t, r)}{P(T = t)} \log_2 \left(\frac{p(t, r)}{P(T = t)} \right) \tag{7}$$

Equivalently, to answer question (c), the conditional entropy for a given property set r is:

$$H(T|R=r) = -\sum_{t \in T} P(T=t|R=r) \log_2 \left(P(T=t|R=r)\right) \tag{8}$$

$$= -\sum_{t \in T} \frac{p(t,r)}{P(R=r)} \log_2 \left(\frac{p(t,r)}{P(R=r)}\right) \tag{9}$$

These conditional entropies are fixed to one particular set of types t or set of properties r. As we are interested in a global insight of a large scale dataset like the LOD cloud, it is not feasible to look at all the individual observations. Rather we need an aggregated value.

Expected Conditional Entropy: For the conditional entropy, there is the possibility of computing the expected conditional entropy $H(R|T)$ which also takes into consideration the probability to actually observe a certain set of types t. The definition of this aggregation is:

$$H(R|T) = \sum_{t \in T} P(T=t) H(R|T=t) \tag{10}$$

$$= -\sum_{t \in T} \sum_{r \in R} p(t,r) \log_2 \left(\frac{p(t,r)}{P(T=t)}\right) \tag{11}$$

And equivalently $H(T|R)$ is for a given set of properties r:

$$H(T|R) = -\sum_{r \in R} \sum_{t \in T} p(t,r) \log_2 \left(\frac{p(t,r)}{P(R=r)}\right) \tag{12}$$

Joint Entropy Finally, we will also take a look at the joint entropy of T and R, which is defined as:

$$H(T,R) = -\sum_{t \in T} \sum_{r \in R} p(t,r) \log_2 \left(p(t,r)\right) \tag{13}$$

Mutual Information To finally answer the question of (d) how far one of the schema information (either properties or types) can explain the respective other, we employ mutual information (MI) [3]. MI is a metric to capture the joint information conveyed by two random variables – and thereby their redundancy. The MI of explicit and implicit schema information of the LOD cloud is defined as:

$$I(T,R) = \sum_{r \in R} \sum_{t \in T} p(t,r) \log_2 \frac{p(t,r)}{P(T=t) \cdot P(R=r)} \tag{14}$$

The log expression in this sum, i.e., the expression $\log_2 \frac{p(t,r)}{P(T=t)\cdot P(R=r)}$ is also known as *pointwise mutual information* (PMI). PMI can be explained as the strength of the correlation of two events, in our case how strongly a given type set and a given property set are associated with each other.

One characteristics of MI is the open range of its values. A normalization of MI to the interval $[-1, 1]$ is given in [13] and involves the entropy of the marginal distributions of T and R. It is defined as

$$I_0(T,R) = \frac{I(T,R)}{\min\left(H(T), H(R)\right)} \tag{15}$$

4.3 Estimating Probabilities from a SchemEX Index

The information SchemEX [7] provides about typeclusters corresponds to the above notion of types sets in TS. The equivalence classes in SchemEX subdivide the typeclusters and are defined by the set of properties the triples have as well as the typecluster the object of triple lies in. Hence, they are more finegrained than the property sets we are interested in. However, if we aggregate the equivalence classes defined by the same set of properties over all attached typeclusters, we obtain exactly the property sets PS, we have introduced above. In this way we can easily construct the set PS from a SchemEX index.

It is important to notice that each entry in the SchemEX index refers to a distinct set of resources. Even if some of the resources are actually located in the same data source. This is provided by the pairwise disjoint character of equivalence classes. In conclusion, we can treat each entry in the index as a different set of resources, even if it is actually reflected by the same URL denoting a common data source.

If we denote with $\mathrm{DS}(t,r)$ the set of entries in the SchemEX index that correspond to the resources with types t and properties r, we can estimate the above probability of observing a resource to have a particular type and property set by:

$$\hat{p}(t,r) = \frac{|\mathrm{DS}(t,r)|}{N}$$

Where N is the number of all data sources used to build the SchemEX and $|\mathrm{DS}(t,r)|$ is the number of data sources containing resources with the type set t and the property set r.

5 Empiric Analysis of Linked Open Data

In the previous section, we have elaborated the metrics to obtain the relevant insights into the information and redundancy encoded in a LOD schema. Furthermore, we also provided an approach to harvest the data about the distribution from a SchemEX index structure. We will now apply this approach to real world data and conduct our analyses. To this end, we need to conduct several steps: obtain datasets (Section 5.1), compute the SchemEX index (Section 5.2), and compute the entropies and mutual information on the schema (Section 5.3). Subsequently, we report and interpret the results.

Table 1. Schema information obtained from the SchemEX indices over our datasets

	TimBL	**TimBL**	**BTC-500M**
Schema construction	stream-based	lossless	stream-based
Size dataset (No. of triples)	11 Mio	11 Mio	500 Mio
No. of type sets	2757	2763	108682
No. of property sets	3838	3753	64520

5.1 Datasets

We use two different datasets for our empiric analysis. As in previous work [8], we use a smaller data set to compute the schema once with our stream-based approach and once in lossless approach. This allows for evaluating the impact of the stream-based schema computation on the employed metrics. This smaller dataset has been crawled with LDSpider starting at Tim Berner-Lee's FOAF file. It contains 11 million triples and – given the starting point of the crawl – we will refer to it as the *TimBL* dataset.[6]

As second dataset, we employ the data provided for the Billion Triples Challenge from 2011.[7] Also this dataset has been crawled from the web and, thus, has similar characteristics. The full BTC 2011 data contains about 2.12 billion triples. While the computation of the schema on this data is not a problem (thanks to the efficient stream-based approach briefly described in Section 3), it is not feasible to determine the joint distribution of type and property combinations on the schema of the full dataset. The reason behind this is the quadratic complexity arising from the computation of the joint distribution and metrics like MI. Thus, we have extracted from the BTC2011 dataset the first 500 million triple for further computations. In the course of this paper, we will refer to this extract as the BTC-500M dataset.

5.2 Computation of the SchemEX Index

In this step, we could leverage the infrastructure setup for our work in [7]. The schema for the TimBL dataset as well for the full BTC2011 dataset have been previously computed and are available online. As mentioned above, we were restricted to run our analytics on a subset of the BTC2011 data. We used the same settings as in [7], using a window size of 50,000 instances for schema extraction. Table 1 gives an overview of the number of type and property combinations observed in the data.

The observed deviations in the number of type sets and property sets in the TimBL dataset are due to the lossless or stream-based schema construction. For details on the computation of the schema, please refer to the literature [8].

5.3 Computation of Entropy and MI

The computation of the relevant metrics is straight forward. The estimates for the probabilities $p(t, r)$ above are central to all the metrics and effectively need only to be ag-

[6] Available from: http://west.uni-koblenz.de/schemex
[7] Available from: http://km.aifb.kit.edu/projects/btc-2011/

gregated and normalized accordingly. However, the number of observed type sets and property sets indicates the number of possible combinations (i.e., $|TS| \times |PS|$). The pragmatic solution to this quadratic development of combinations is not to compute all of the probabilities, but only those which actually have a non zero value. This does not affect the results of the computed metric, as zero probabilities do not affect their overall values.

6 Results of our Analysis

Table 2. Overview: Entropy and MI in the schemata of the considered datasets

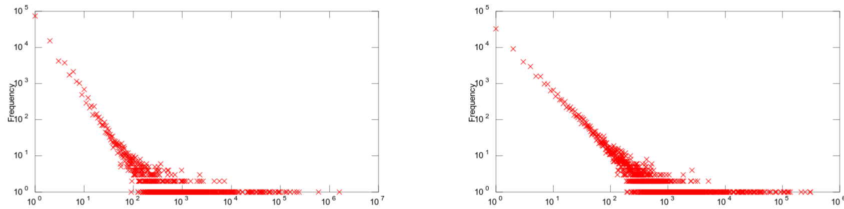
		TimBL	**TimBL**	**BTC-500M**
Schema construction		stream-based	lossless	stream-based
Entropy of type sets	$H(T)$	4.8168	5.0237	6.4941
Normalized entropy of type sets	$H_0(T)$	0.4215	0.4394	0.3882
Entropy of property sets	$H(R)$	5.9574	6.3533	7.6317
Normalized entropy of property sets	$H_0(R)$	0.5004	0.5351	0.4777
Conditional entropy, given properties	$H(T\|R)$	0.3855	0.3819	0.8225
Conditional entropy, given types	$H(R\|T)$	2.3716	2.3983	2.7305
Joint entropy	$H(T, R)$	6.3429	6.7352	8.4554
Mutual Information	$I(T, R)$	3.5858	3.9550	4.9024
Normalized MI	$I_0(T, R)$	0.8026	0.7873	0.7549

Table 2 gives an overview of the computed metrics on the evaluated datasets. We will now go into details about the single figures.

Entropy in type and property sets: We can observe on both datasets, that the property sets convey more information than type sets. This fact is reflected also in the normalized entropy and is therefore independent of the actual number of type sets and property sets. This corresponds to the property sets being more uniformly distributed over resources. In Figure 2, we have plotted the distribution of data sources into type sets (a) and property sets (b).

Conditional Entropies: Looking at the conditional entropies reveals some interesting insights. Recall that the aggregation we chose for the conditional entropy provides us with the expected entropy, given a certain type set or property set. We can see in Table 2 that the entropy given a property set is far lower than given a type set. In conclusion: knowing the properties of a resource already tells us a lot about the type of the resource, as the entropy of the conditional distribution can be expected to be quite low. On the contrary, when knowing the type of a resource the entropy of the distribution of the property sets can be expected to be still relatively high (when compared to the entropy of the marginal distribution).

Mutual Information: Finally, the value of the normalized MI gives us insights how much one information (either properties or types) explains the respective other. On



(a) (b)

Figure 2. Distributions of data sources in type sets and property sets in the BTC-500M dataset.

the larger BTC-500M dataset, we observe a normalized MI of 75.49%. Accordingly, extracting only type or only property information from LOD can already explain a quite large share of the contained information. However, given our observations a significant part of the schema information is encoded also in the respective other part.

Regarding PMI, we have analyzed which values occur for the observed type sets and property sets. Figure 3 shows a plot of how the normalized PMI values are distributed. This shows that in general we can observe a positive correlation between the observed[8] type sets and property sets. For many combinations, we can even see that the correlation is quite high. We can also observe some spikes which indicate that some levels of correlation are more typical than others.

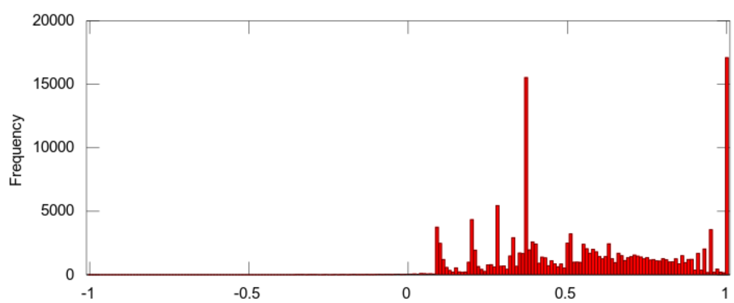


Figure 3. Distributions of PMI values.

7 Discussion of the Results

Before interpreting our results, we need to look at the comparison of the stream-based and the lossless schema construction on the TimBL dataset. As in previous work, we

[8] We omitted the calculation of unobserved combinations which would technically amount for a high peak at -1 but does not really provide us with the insights we are seeking here.

can see that even though the stream-based schema extraction does make some mistake, the deviations are minimal. Given that the stream-based approach allows us to analyze a much larger dataset and, thus, to have a more reliable sample for estimating the joint distribution of type sets and property sets. The small tradeoff in accuracy of the computed stream-based vs. lossless schema is justified.

The observations we have made on the larger BTC-500M dataset provide us with some interesting insights into the form and structure of schema information on the LOD cloud. First of all, the distribution of type sets and property sets have a relatively high entropy. We can conclude that the structure of the data is not dominated by a few combinations of types or properties. Accordingly for the extraction of schematic information, we cannot reduce the schema to a small and fixed structure but need to consider the wide variety of type and property information. Otherwise the schema would loose too much information.

A second observation is the dependency between types and properties. The conditional entropy reveals that the properties of a resource tell much more about its type than the other way around. This observation is interesting for various applications. For instance, suggesting a data engineer the types of a resource based on the already modeled properties seems quite promising. However, suggesting or predicting the set of properties for a given type is a much harder task. We assume that this observation can also be seen as an evidence that property information on the LOD cloud actually considers implicit or explicit agreements about the domain and range of the according property.

The observation on the conditional entropies is interesting in particular also with the measured MI in the schema. As type set or property set information can explain for about 75% of the total information in the schema, this motivates the potential of building a schema only over one of these two types. Given further, that the properties show stronger implications for the types, a schema based solely on property information seems a reasonable approach. However, this approach entails a significant loss of schema information and its applicability depends on the concrete scenario.

8 Conclusions and Future Work

In this paper, we have conducted an in depth analysis of schema information on Linked Open Data. In particular, we have addressed the question of correlation between the types of resources and their properties. Based on the schema derived from 11 Million triples of a dataset crawled starting with TB Lee's FOAF profile and 500 Million triples taken from the BTC 2011 challenge, we have computed various entropy metrics as well as mutual information. In conclusion, we observe a reasonably high correlation between the types and properties attached to resources. As more detailed conclusion, we can derive that the properties of a resource are rather indicative for the type the resource. In the other direction, the indication is less strong.

As future work, we plan to deepen these insights and incorporate the obtained deeper understanding into various applications. Therefore, we will look into the details of the conditional distributions for given type sets and property sets. In this way, we might identify which sets of types and properties allow for highly precise predictions of the respective other schema information. On the application side, we plan to use the gained insights for various purposes: index compression and index correction for SchemEX as well as the detection of schema patterns that are stable enough – and thereby suitable – for constructing an API for accessing LOD resources.

Acknowledgements

We thank Andreas Harth for providing the crawled TimBL dataset that was used in our experiment. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 257859, ROBUST and grant agreement no. 287975, SocialSensor.

References

1. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets with the void vocabulary. http://www.w3.org/TR/void/, last visited 22.07.2012
2. Auer, S., Demter, J., Martin, M., Lehman, J.: Lodstats - an extensible framework for high-performance dataset analytics. In: EKAW (2012), (to appear)
3. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience (1991)
4. Harth, A., Hose, K., Karnstedt, M., Polleres, A., Sattler, K.U., Umbrich, J.: Data summaries for on-demand queries over linked data. In: WWW. pp. 411–420. ACM (2010)
5. Heath, T., Bizer, C.: Linked Data: Evolving the Web Into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool (2011)
6. Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., Decker, S.: An empirical survey of linked data conformance. Journal of Web Semantics (to appear) (2012)
7. Konrath, M., Gottron, T., Scherp, A.: Schemex – web-scale indexed schema extraction of linked open data. In: Semantic Web Challenge, Submission to the Billion Triple Track (2011)
8. Konrath, M., Gottron, T., Staab, S., Scherp, A.: Schemex—efficient construction of a data catalogue by stream-based indexing of linked data. Journal of Web Semantics (2012), (to appear)
9. Maduko, A., Anyanwu, K., Sheth, A., Schliekelman, P.: Graph summaries for subgraph frequency estimation. In: Proceedings of the 5th European Semantic Web Conference. LNCS, Springer Verlag, Berlin, Heidelberg (June 2008)
10. Neumann, T., Moerkotte, G.: Characteristic sets: Accurate cardinality estimation for rdf queries with multiple joins. In: Proceedings of the 27th International Conference on Data Engineering, ICDE 2011, April 11-16, 2011, Hannover, Germany. pp. 984–994 (2011)
11. Neumann, T., Weikum, G.: Scalable join processing on very large rdf graphs. In: SIGMOD Conference. pp. 627–640. ACM (2009)
12. Shannon, C.: A mathematical theory of communication. Bell System Technical Journal 27, 379–423 and 623–656 (July and October 1948)
13. Yao, Y.Y.: Information-theoretic measures for knowledge discovery and data mining. In: Karmeshu, J.K. (ed.) Entropy Measures, Maximum Entropy Principle and Emerging Applications, pp. 115–136. Springer, Berlin (2003)

Bisher erschienen

Arbeitsberichte aus dem Fachbereich Informatik

(http://www.uni-koblenz-landau.de/koblenz/fb4/publications/Reports/arbeitsberichte)

Ansgar Scherp, Thomas Gottron, Malte Knauf, Stefan Scheglmann, Explicit and Implicit Schema Information on the Linked Open Data Cloud: Joined Forces or Antagonists? Arbeitsberichte aus dem Fachbereich Informatik 6/2012

Harald von Kortzfleisch, Ilias Mokanis, Dorothée Zerwas, Introducing Entrepreneurial Design Thinking, Arbeitsberichte aus dem Fachbereich Informatik 5/2012

Ansgar Scherp, Daniel Eißing, Carsten Saathoff, Integrating Multimedia Metadata Standarts and Metadata Formats with the Multimedia Metadata Ontology: Method and Examples, Arbeitsberichte aus dem Fachbereich Informatik 4/2012

Martin Surrey,Björn Lilge, Ludwig Paulsen, Marco Wolf, Markus Aldenhövel, Mike Reuthel, Roland Diehl, Integration von CRM-Systemen mit Kollaborations-Systemen am Beispiel von DocHouse und Lotus Quickr, Arbeitsberichte aus dem Fachbereich Informatik 3/2012

Martin Surrey, Roland Diehl, DOCHOUSE: Opportunity Management im Partnerkanal (IBM Lotus Quickr), Arbeitsberichte aus dem Fachbereich Informatik 2/2012

Mark Schneider, Ansgar Scherp, Comparing a Grid-based vs. List-based Approach for Faceted Search of Social Media Data on Mobile Devices, Arbeitsberichte aus dem Fachbereich Informatik 1/2012

Petra Schubert, Femi Adisa, Cloud Computing for Standard ERP Systems: Reference Framework and Research Agenda, Arbeitsberichte aus dem Fachbereich Informatik 16/2011

Oleg V. Kryuchin, Alexander A. Arzamastsev, Klaus G. Troitzsch, Natalia A. Zenkova, Simulating social objects with an artificial network using a computer cluster, Arbeitsberichte aus dem Fachbereich Informatik 15/2011

Oleg V. Kryuchin, Alexander A. Arzamastsev, Klaus G. Troitzsch, Simulating medical objects using an artificial network whose structure is based on adaptive resonance theory, Arbeitsberichte aus dem Fachbereich Informatik 14/2011

Oleg V. Kryuchin, Alexander A. Arzamastsev, Klaus G. Troitzsch, Comparing the efficiency of serial and parallel algorithms for training artificial neural networks using computer clusters, Arbeitsberichte aus dem Fachbereich Informatik, 13/2011

Oleg V. Kryuchin, Alexander A. Arzamastsev, Klaus G. Troitzsch, A parallel algorithm for selecting activation functions of an artificial network, Arbeitsberichte aus dem Fachbereich Informatik 12/2011

Katharina Bräunlich, Rüdiger Grimm, Andreas Kasten, Sven Vowé, Nico Jahn, Der neue Personalausweis zur Authentifizierung von Wählern bei Onlinewahlen, Arbeitsberichte aus dem Fachbereich Informatik 11/2011

Daniel Eißing, Ansgar Scherp, Steffen Staab, Formal Integration of Individual Knowledge Work and Organizational Knowledge Work with the Core Ontology *strukt*, Arbeitsberichte aus dem Fachbereich Informatik 10/2011

Bernhard Reinert, Martin Schumann, Stefan Müller, Combined Non-Linear Pose Estimation from Points and Lines, Arbeitsberichte aus dem Fachbereich Informatik 9/2011

Tina Walber, Ansgar Scherp, Steffen Staab, Towards the Understanding of Image Semantics by Gaze-based Tag-to-Region Assignments, Arbeitsberichte aus dem Fachbereich Informatik 8/2011

Alexander Kleinen, Ansgar Scherp, Steffen Staab, Mobile Facets – Faceted Search and Exploration of Open Social Media Data on a Touchscreen Mobile Phone, Arbeitsberichte aus dem Fachbereich Informatik 7/2011

Anna Lantsberg, Klaus G. Troitzsch, Towards A Methodology of Developing Models of E-Service Quality Assessment in Healthcare, Arbeitsberichte aus dem Fachbereich Informatik 6/2011

Ansgar Scherp, Carsten Saathoff, Thomas Franz, Steffen Staab, Designing Core Ontologies, Arbeitsberichte aus dem Fachbereich Informatik 5/2011

Oleg V. Kryuchin, Alexander A. Arzamastsev, Klaus G. Troitzsch, The prediction of currency exchange rates using artificial neural networks, Arbeitsberichte aus dem Fachbereich Informatik 4/2011

Klaus G. Troitzsch, Anna Lantsberg, Requirements for Health Care Related Websites in Russia: Results from an Analysis of American, British and German Examples, Arbeitsberichte aus dem Fachbereich Informatik 3/2011

Klaus G. Troitzsch, Oleg Kryuchin, Alexander Arzamastsev, A universal simulator based on artificial neural networks for computer clusters, Arbeitsberichte aus dem Fachbereich Informatik 2/2011

Klaus G. Troitzsch, Natalia Zenkova, Alexander Arzamastsev, Development of a technology of designing intelligent information systems for the estimation of social objects, Arbeitsberichte aus dem Fachbereich Informatik 1/2011

Kurt Lautenbach, A Petri Net Approach for Propagating Probabilities and Mass Functions, Arbeitsberichte aus dem Fachbereich Informatik 13/2010

Claudia Schon, Linkless Normal Form for ALC Concepts, Arbeitsberichte aus dem Fachbereich Informatik 12/2010

Alexander Hug, Informatik hautnah erleben, Arbeitsberichte aus dem Fachbereich Informatik 11/2010

Marc Santos, Harald F.O. von Kortzfleisch, Shared Annotation Model – Ein Datenmodell für kollaborative Annotationen, Arbeitsberichte aus dem Fachbereich Informatik 10/2010

Gerd Gröner, Steffen Staab, Categorization and Recognition of Ontology Refactoring Pattern, Arbeitsberichte aus dem Fachbereich Informatik 9/2010

Daniel Eißing, Ansgar Scherp, Carsten Saathoff, Integration of Existing Multimedia Metadata Formats and Metadata Standards in the M3O, Arbeitsberichte aus dem Fachbereich Informatik 8/2010

Stefan Scheglmann, Ansgar Scherp, Steffen Staab, Model-driven Generation of APIs for OWL-based Ontologies, Arbeitsberichte aus dem Fachbereich Informatik 7/2010

Daniel Schmeiß, Ansgar Scherp, Steffen Staab, Integrated Mobile Visualization and Interaction of Events and POIs, Arbeitsberichte aus dem Fachbereich Informatik 6/2010

Rüdiger Grimm, Daniel Pähler, E-Mail-Forensik – IP-Adressen und ihre Zuordnung zu Internet-Teilnehmern und ihren Standorten, Arbeitsberichte aus dem Fachbereich Informatik 5/2010

Christoph Ringelstein, Steffen Staab, PAPEL: Syntax and Semantics for Provenance-Aware Policy Definition, Arbeitsberichte aus dem Fachbereich Informatik 4/2010

Nadine Lindermann, Sylvia Valcárcel, Harald F.O. von Kortzfleisch, Ein Stufenmodell für kollaborative offene Innovationsprozesse in Netzwerken kleiner und mittlerer Unternehmen mit Web 2.0, Arbeitsberichte aus dem Fachbereich Informatik 3/2010

Maria Wimmer, Dagmar Lück-Schneider, Uwe Brinkhoff, Erich Schweighofer, Siegfried Kaiser, Andreas Wieber, Fachtagung Verwaltungsinformatik FTVI Fachtagung Rechtsinformatik FTRI 2010, Arbeitsberichte aus dem Fachbereich Informatik 2/2010

Max Braun, Ansgar Scherp, Steffen Staab, Collaborative Creation of Semantic Points of Interest as Linked Data on the Mobile Phone, Arbeitsberichte aus dem Fachbereich Informatik 1/2010

Marc Santos, Einsatz von „Shared In-situ Problem Solving“ Annotationen in kollaborativen Lern- und Arbeitsszenarien, Arbeitsberichte aus dem Fachbereich Informatik 20/2009

Carsten Saathoff, Ansgar Scherp, Unlocking the Semantics of Multimedia Presentations in the Web with the Multimedia Metadata Ontology, Arbeitsberichte aus dem Fachbereich Informatik 19/2009

Christoph Kahle, Mario Schaarschmidt, Harald F.O. von Kortzfleisch, Open Innovation: Kundenintegration am Beispiel von IPTV, Arbeitsberichte aus dem Fachbereich Informatik 18/2009

Dietrich Paulus, Lutz Priese, Peter Decker, Frank Schmitt, Pose-Tracking Forschungsbericht, Arbeitsberichte aus dem Fachbereich Informatik 17/2009

Andreas Fuhr, Tassilo Horn, Andreas Winter, Model-Driven Software Migration Extending SOMA, Arbeitsberichte aus dem Fachbereich Informatik 16/2009

Eckhard Großmann, Sascha Strauß, Tassilo Horn, Volker Riediger, Abbildung von grUML nach XSD soamig, Arbeitsberichte aus dem Fachbereich Informatik 15/2009

Kerstin Falkowski, Jürgen Ebert, The STOR Component System Interim Report, Arbeitsberichte aus dem Fachbereicht Informatik 14/2009

Sebastian Magnus, Markus Maron, An Empirical Study to Evaluate the Location of Advertisement Panels by Using a Mobile Marketing Tool, Arbeitsberichte aus dem Fachbereich Informatik 13/2009

Sebastian Magnus, Markus Maron, Konzept einer Public Key Infrastruktur in iCity, Arbeitsberichte aus dem Fachbereich Informatik 12/2009

Sebastian Magnus, Markus Maron, A Public Key Infrastructure in Ambient Information and Transaction Systems, Arbeitsberichte aus dem Fachbereich Informatik 11/2009

Ammar Mohammed, Ulrich Furbach, Multi-agent systems: Modeling and Virification using Hybrid Automata, Arbeitsberichte aus dem Fachbereich Informatik 10/2009

Andreas Sprotte, Performance Measurement auf der Basis von Kennzahlen aus betrieblichen Anwendungssystemen: Entwurf eines kennzahlengestützten Informationssystems für einen Logistikdienstleister, Arbeitsberichte aus dem Fachbereich Informatik 9/2009

Gwendolin Garbe, Tobias Hausen, Process Commodities: Entwicklung eines Reifegradmodells als Basis für Outsourcingentscheidungen, Arbeitsberichte aus dem Fachbereich Informatik 8/2009

Petra Schubert et. al., Open-Source-Software für das Enterprise Resource Planning, Arbeitsberichte aus dem Fachbereich Informatik 7/2009

Ammar Mohammed, Frieder Stolzenburg, Using Constraint Logic Programming for Modeling and Verifying Hierarchical Hybrid Automata, Arbeitsberichte aus dem Fachbereich Informatik 6/2009

Tobias Kippert, Anastasia Meletiadou, Rüdiger Grimm, Entwurf eines Common Criteria-Schutzprofils für Router zur Abwehr von Online-Überwachung, Arbeitsberichte aus dem Fachbereich Informatik 5/2009

Hannes Schwarz, Jürgen Ebert, Andreas Winter, Graph-based Traceability – A Comprehensive Approach. Arbeitsberichte aus dem Fachbereich Informatik 4/2009

Anastasia Meletiadou, Simone Müller, Rüdiger Grimm, Anforderungsanalyse für Risk-Management-Informationssysteme (RMIS), Arbeitsberichte aus dem Fachbereich Informatik 3/2009

Ansgar Scherp, Thomas Franz, Carsten Saathoff, Steffen Staab, A Model of Events based on a Foundational Ontology, Arbeitsberichte aus dem Fachbereich Informatik 2/2009

Frank Bohdanovicz, Harald Dickel, Christoph Steigner, Avoidance of Routing Loops, Arbeitsberichte aus dem Fachbereich Informatik 1/2009

Stefan Ameling, Stephan Wirth, Dietrich Paulus, Methods for Polyp Detection in Colonoscopy Videos: A Review, Arbeitsberichte aus dem Fachbereich Informatik 14/2008

Tassilo Horn, Jürgen Ebert, Ein Referenzschema für die Sprachen der IEC 61131-3, Arbeitsberichte aus dem Fachbereich Informatik 13/2008

Thomas Franz, Ansgar Scherp, Steffen Staab, Does a Semantic Web Facilitate Your Daily Tasks?, Arbeitsberichte aus dem Fachbereich Informatik 12/2008

Norbert Frick, Künftige Anfordeungen an ERP-Systeme: Deutsche Anbieter im Fokus, Arbeitsberichte aus dem Fachbereicht Informatik 11/2008

Jürgen Ebert, Rüdiger Grimm, Alexander Hug, Lehramtsbezogene Bachelor- und Masterstudiengänge im Fach Informatik an der Universität Koblenz-Landau, Campus Koblenz, Arbeitsberichte aus dem Fachbereich Informatik 10/2008

Mario Schaarschmidt, Harald von Kortzfleisch, Social Networking Platforms as Creativity Fostering Systems: Research Model and Exploratory Study, Arbeitsberichte aus dem Fachbereich Informatik 9/2008

Bernhard Schueler, Sergej Sizov, Steffen Staab, Querying for Meta Knowledge, Arbeitsberichte aus dem Fachbereich Informatik 8/2008

Stefan Stein, Entwicklung einer Architektur für komplexe kontextbezogene Dienste im mobilen Umfeld, Arbeitsberichte aus dem Fachbereich Informatik 7/2008

Matthias Bohnen, Lina Brühl, Sebastian Bzdak, RoboCup 2008 Mixed Reality League Team Description, Arbeitsberichte aus dem Fachbereich Informatik 6/2008

Bernhard Beckert, Reiner Hähnle, Tests and Proofs: Papers Presented at the Second International Conference, TAP 2008, Prato, Italy, April 2008, Arbeitsberichte aus dem Fachbereich Informatik 5/2008

Klaas Dellschaft, Steffen Staab, Unterstützung und Dokumentation kollaborativer Entwurfs- und Entscheidungsprozesse, Arbeitsberichte aus dem Fachbereich Informatik 4/2008

Rüdiger Grimm: IT-Sicherheitsmodelle, Arbeitsberichte aus dem Fachbereich Informatik 3/2008

Rüdiger Grimm, Helge Hundacker, Anastasia Meletiadou: Anwendungsbeispiele für Kryptographie, Arbeitsberichte aus dem Fachbereich Informatik 2/2008

Markus Maron, Kevin Read, Michael Schulze: CAMPUS NEWS – Artificial Intelligence Methods Combined for an Intelligent Information Network, Arbeitsberichte aus dem Fachbereich Informatik 1/2008

Lutz Priese,Frank Schmitt, Patrick Sturm, Haojun Wang: BMBF-Verbundprojekt 3D-RETISEG Abschlussbericht des Labors Bilderkennen der Universität Koblenz-Landau, Arbeitsberichte aus dem Fachbereich Informatik 26/2007

Stephan Philippi, Alexander Pinl: Proceedings 14. Workshop 20.-21. September 2007 Algorithmen und Werkzeuge für Petrinetze, Arbeitsberichte aus dem Fachbereich Informatik 25/2007

Ulrich Furbach, Markus Maron, Kevin Read: CAMPUS NEWS – an Intelligent Bluetooth-based Mobile Information Network, Arbeitsberichte aus dem Fachbereich Informatik 24/2007

Ulrich Furbach, Markus Maron, Kevin Read: CAMPUS NEWS - an Information Network for Pervasive Universities, Arbeitsberichte aus dem Fachbereich Informatik 23/2007

Lutz Priese: Finite Automata on Unranked and Unordered DAGs Extented Version, Arbeitsberichte aus dem Fachbereich Informatik 22/2007

Mario Schaarschmidt, Harald F.O. von Kortzfleisch: Modularität als alternative Technologie- und Innovationsstrategie, Arbeitsberichte aus dem Fachbereich Informatik 21/2007

Kurt Lautenbach, Alexander Pinl: Probability Propagation Nets, Arbeitsberichte aus dem Fachbereich Informatik 20/2007

Rüdiger Grimm, Farid Mehr, Anastasia Meletiadou, Daniel Pähler, Ilka Uerz: SOA-Security, Arbeitsberichte aus dem Fachbereich Informatik 19/2007

Christoph Wernhard: Tableaux Between Proving, Projection and Compilation, Arbeitsberichte aus dem Fachbereich Informatik 18/2007

Ulrich Furbach, Claudia Obermaier: Knowledge Compilation for Description Logics, Arbeitsberichte aus dem Fachbereich Informatik 17/2007

Fernando Silva Parreiras, Steffen Staab, Andreas Winter: TwoUse: Integrating UML Models and OWL Ontologies, Arbeitsberichte aus dem Fachbereich Informatik 16/2007

Rüdiger Grimm, Anastasia Meletiadou: Rollenbasierte Zugriffskontrolle (RBAC) im Gesundheitswesen, Arbeitsberichte aud dem Fachbereich Informatik 15/2007

Ulrich Furbach, Jan Murray, Falk Schmidsberger, Frieder Stolzenburg: Hybrid Multiagent Systems with Timed Synchronization-Specification and Model Checking, Arbeitsberichte aus dem Fachbereich Informatik 14/2007

Björn Pelzer, Christoph Wernhard: System Description:“E-KRHyper“, Arbeitsberichte aus dem Fachbereich Informatik, 13/2007

Ulrich Furbach, Peter Baumgartner, Björn Pelzer: Hyper Tableaux with Equality, Arbeitsberichte aus dem Fachbereich Informatik, 12/2007

Ulrich Furbach, Markus Maron, Kevin Read: Location based Informationsystems, Arbeitsberichte aus dem Fachbereich Informatik, 11/2007

Philipp Schaer, Marco Thum: State-of-the-Art: Interaktion in erweiterten Realitäten, Arbeitsberichte aus dem Fachbereich Informatik, 10/2007

Ulrich Furbach, Claudia Obermaier: Applications of Automated Reasoning, Arbeitsberichte aus dem Fachbereich Informatik, 9/2007

Jürgen Ebert, Kerstin Falkowski: A First Proposal for an Overall Structure of an Enhanced Reality Framework, Arbeitsberichte aus dem Fachbereich Informatik, 8/2007

Lutz Priese, Frank Schmitt, Paul Lemke: Automatische See-Through Kalibrierung, Arbeitsberichte aus dem Fachbereich Informatik, 7/2007

Rüdiger Grimm, Robert Krimmer, Nils Meißner, Kai Reinhard, Melanie Volkamer, Marcel Weinand, Jörg Helbach: Security Requirements for Non-political Internet Voting, Arbeitsberichte aus dem Fachbereich Informatik, 6/2007

Daniel Bildhauer, Volker Riediger, Hannes Schwarz, Sascha Strauß, „grUML – Eine UML-basierte Modellierungssprache für T-Graphen“, Arbeitsberichte aus dem Fachbereich Informatik, 5/2007

Richard Arndt, Steffen Staab, Raphaël Troncy, Lynda Hardman: Adding Formal Semantics to MPEG-7: Designing a Well Founded Multimedia Ontology for the Web, Arbeitsberichte aus dem Fachbereich Informatik, 4/2007

Simon Schenk, Steffen Staab: Networked RDF Graphs, Arbeitsberichte aus dem Fachbereich Informatik, 3/2007

Rüdiger Grimm, Helge Hundacker, Anastasia Meletiadou: Anwendungsbeispiele für Kryptographie, Arbeitsberichte aus dem Fachbereich Informatik, 2/2007

Anastasia Meletiadou, J. Felix Hampe: Begriffsbestimmung und erwartete Trends im IT-Risk-Management, Arbeitsberichte aus dem Fachbereich Informatik, 1/2007

„Gelbe Reihe“
(http://www.uni-koblenz.de/fb4/publikationen/gelbereihe)

Lutz Priese: Some Examples of Semi-rational and Non-semi-rational DAG Languages. Extended Version, Fachberichte Informatik 3-2006

Kurt Lautenbach, Stephan Philippi, and Alexander Pinl: Bayesian Networks and Petri Nets, Fachberichte Informatik 2-2006

Rainer Gimnich and Andreas Winter: Workshop Software-Reengineering und Services, Fachberichte Informatik 1-2006

Kurt Lautenbach and Alexander Pinl: Probability Propagation in Petri Nets, Fachberichte Informatik 16-2005

Rainer Gimnich, Uwe Kaiser, and Andreas Winter: 2. Workshop ''Reengineering Prozesse'' – Software Migration, Fachberichte Informatik 15-2005

Jan Murray, Frieder Stolzenburg, and Toshiaki Arai: Hybrid State Machines with Timed Synchronization for Multi-Robot System Specification, Fachberichte Informatik 14-2005

Reinhold Letz: FTP 2005 – Fifth International Workshop on First-Order Theorem Proving, Fachberichte Informatik 13-2005

Bernhard Beckert: TABLEAUX 2005 – Position Papers and Tutorial Descriptions, Fachberichte Informatik 12-2005

Dietrich Paulus and Detlev Droege: Mixed-reality as a challenge to image understanding and artificial intelligence, Fachberichte Informatik 11-2005

Jürgen Sauer: 19. Workshop Planen, Scheduling und Konfigurieren / Entwerfen, Fachberichte Informatik 10-2005

Pascal Hitzler, Carsten Lutz, and Gerd Stumme: Foundational Aspects of Ontologies, Fachberichte Informatik 9-2005

Joachim Baumeister and Dietmar Seipel: Knowledge Engineering and Software Engineering, Fachberichte Informatik 8-2005

Benno Stein and Sven Meier zu Eißen: Proceedings of the Second International Workshop on Text-Based Information Retrieval, Fachberichte Informatik 7-2005

Andreas Winter and Jürgen Ebert: Metamodel-driven Service Interoperability, Fachberichte Informatik 6-2005

Joschka Boedecker, Norbert Michael Mayer, Masaki Ogino, Rodrigo da Silva Guerra, Masaaki Kikuchi, and Minoru Asada: Getting closer: How Simulation and Humanoid League can benefit from each other, Fachberichte Informatik 5-2005

Torsten Gipp and Jürgen Ebert: Web Engineering does profit from a Functional Approach, Fachberichte Informatik 4-2005

Oliver Obst, Anita Maas, and Joschka Boedecker: HTN Planning for Flexible Coordination Of Multiagent Team Behavior, Fachberichte Informatik 3-2005

Andreas von Hessling, Thomas Kleemann, and Alex Sinner: Semantic User Profiles and their Applications in a Mobile Environment, Fachberichte Informatik 2-2005

Heni Ben Amor and Achim Rettinger: Intelligent Exploration for Genetic Algorithms – Using Self-Organizing Maps in Evolutionary Computation, Fachberichte Informatik 1-2005