Mining Significant Maximum Cardinalities in Knowledge Bases

Arnaud Giacometti, Béatrice Markhoff, and Arnaud Soulet(✉)

Université de Tours, LIFAT, Blois, France
{arnaud.giacometti,beatrice.markhoff,arnaud.soulet}@univ-tours.fr

Abstract. Semantic Web connects huge knowledge bases whose content has been generated from collaborative platforms and by integration of heterogeneous databases. Naturally, these knowledge bases are incomplete and contain erroneous data. Knowing their data quality is an essential long-term goal to guarantee that querying them returns reliable results. Having cardinality constraints for roles would be an important advance to distinguish correctly and completely described individuals from those having data either incorrect or insufficiently informed. In this paper, we propose a method for automatically discovering from the knowledge base's content the maximum cardinality of roles for each concept, when it exists. This method is robust thanks to the use of Hoeffding's inequality. We also design an algorithm, named C3M, for an exhaustive search of such constraints in a knowledge base benefiting from pruning properties that drastically reduce the search space. Experiments conducted on DBpedia demonstrate the scaling up of C3M, and also highlight the robustness of our method, with a precision higher than 95%.

Keywords: Cardinality mining · Contextual constraint · Knowledge base

1 Introduction

With the rise of the Semantic Web, knowledge bases (that we will denote KB) are growing and multiplying. At the worldwide level knowledge hubs are built from collaborative platforms, either by extraction from Wikipedia as DBpedia [1] or collaboratively collecting knowledge as for Wikidata [6], or integrating various sources using information retrieval algorithms as for YAGO [21]. These very large KBs represent a wealth of information for applications, as this is the case with Wikipedia for human beings. On a smaller scale, more and more knowledge bases are published on the Web, built from diverse data sources following Extract-Transform-Load integration processes that are based on a shared ontology (ontology-based data integration).

Due to the way they are generated, all of these KBs need to be enriched with more information to evaluate their quality with respect to the represented reality,

C. Ghidini et al. (Eds.): ISWC 2019, LNCS 11778, pp. 182–199, 2019.
https://doi.org/10.1007/978-3-030-30793-6_11

and reverse engineering techniques have already been considered to automatically obtain useful declarations such as keys [16,19]. In this paper we propose to automatically discover another kind of useful declaration about the represented data in a given KB: *role maximum cardinalities*. In knowledge representation, numerical restrictions which specify the number of occurrences of a role are particularly useful [2]. For example, a numerical restriction can be used to describe a concept[1] C as the set of individuals who have at most 3 children. Moreover, a numerical restriction can be used to declare a *maximum cardinality constraint on the role R in the context C*, for instance on the role `parent` in the context `Person`, for declaring that individuals of concept `Person` have at most twice the role `parent`. Such a declaration allows reasoners to infer whether all the assertions on role R exist in the KB for any individual belonging to C. This can be used to supplement the answers to queries with precise information on their quality in terms of *recall* with respect to reality [20].

Table 1. Cardinality distributions for some contexts/roles in DBpedia (with the role cardinality i, the number of individuals n_i having i times this role, the likelihood τ_i and the pessimistic likelihood $\widetilde{\tau}_i$ that are defined in Sect. 4.1)

`Person` / `birthYear`

i	n_i	τ_i	$\widetilde{\tau}_i$
1	**159,841**	**0.999**	**0.996**
2	91	0.928	0.775
3	4	0.571	0.000
4	2	0.667	0.000
5	1	1.000	0.000

`Person` / `parent`

i	n_i	τ_i	$\widetilde{\tau}_i$
1	10,643	0.529	0.518
2	**9,392**	**0.991**	**0.975**
3	75	0.882	0.718
4	9	0.900	0.420
6	1	1.000	0.000

$\top$ / `team`

i	n_i	τ_i	$\widetilde{\tau}_i$
1	1,221,202	0.901	0.900
2	20,505	0.153	0.148
3	16,876	0.148	0.144
...	...	...	...
20	2	1.000	0.000

`FootballMatch` / `team`

i	n_i	τ_i	$\widetilde{\tau}_i$
1	26	0.008	0.000
2	**3,092**	**0.998**	**0.971**
3	3	0.500	0.000
4	2	0.667	0.000
5	1	1.000	0.000

To the best of our knowledge there is only one work dedicated to the extraction of cardinality constraint from a KB [15], maybe because compared to the traditional database framework, extracting *significant* cardinality constraints from a KB is a far more challenging task. Indeed, we are facing three important challenges. A *first challenge* is that a KB generally contains *inconsistent* data, either because of errors or because of duplicate descriptions. Due to these inconsistencies, the *observed* maximum cardinality for a role in a KB cannot be considered

[1] We use the Description Logics (DL) [2] terminology, as DL are the theoretical foundations of OWL, so we use the terms *concept* (i.e. class), *role* (i.e. property), *individual* and *fact* (i.e. instances).

to be its true maximum cardinality. For example, it is expected that a person will have at most one birth year and two parents. However, considering the roles `birthYear` and `parent` in DBpedia (see Table 1), some persons have 5 birth years or 6 parents. These few inconsistent assertions should not influence the maximum cardinality discovery. Then, a *second challenge* is that a KB is often *incomplete* for a given role. For this reason, the *most frequently observed* cardinality for a role in a KB cannot be considered to be its true maximum cardinality. Typically, most people described in DBpedia have only one informed parent. Nonetheless, we have to take into account that many people have two informed parents for not underestimating the maximum cardinality of the role `parent`. Finally, a *third challenge* is that the expected constraints depend on a *context*. For instance in DBpedia the role `team` is used to inform the teams to which a person has belonged and the teams of a football match. Thus, it is not possible to determine the maximum cardinality of the role `team` in DBpedia (context $\top$), but its maximum cardinality is expected to be 2 in the context of `FootballMatch`. Consequently, instead of exploring each role of a knowledge base, we have to explore each role for each concept. This leads to a huge search space and therefore it is necessary to prune it without missing relevant constraints. But, conversely, we have to avoid extracting redundant constraints. If we identify that a person has at most one birth year, it would be a shame to overwhelm the end user with the cardinality of `birthYear` for artists, scientists and so on.

Taking into account these challenges, we present in this paper two main contributions. Our first contribution is to propose *a method for computing a significant maximum cardinality*. The significance is guaranteed by the use of Hoeffding's inequality for computing corrected likelihood estimates of maximum cardinality. We show with experiments using DBpedia that we extract only reliable maximum cardinalities. More precisely, contrary to [15] it is important to note that we output a maximum cardinality if and only if it is actually significant. Our second contribution is C3M[2], *an algorithm for enumerating the set of all contextual maximum cardinalities* that are minimal (Definition 2) and significant (Definition 4). We use two sound pruning criteria that drastically reduce the exploration space, and ensure the scalability of C3M for large KBs. It is also interesting to notice that we implemented C3M in such a way that it explores Web KBs via their public SPARQL endpoints without centralizing data.

This paper is structured as follows. Section 2 reviews some related works. In Sect. 3, we first introduce some basic notions and formalize the problem. Then, in Sect. 4, we show how to detect a significant maximum cardinality of a role. Next, in Sect. 5, we present our algorithm C3M. Section 6 provides experimental results on DBpedia that shows its efficiency and its scalability, together with the meaningfulness of discovered constraints. We conclude in Sect. 7.

[2] The prototype and the results are available at https://github.com/asoulet/c3m, both in CSV and in RDF (Turtle); we provide also the schema of our constraints expressed in RDF.

2 Related Work

To increase knowledge about the quality of data contained in KB, some proposals calculate quality indicators like completeness [17] or representativeness [18], while others are interested in the enrichment of individuals or concepts with fine-grained assertions or constraints. Our proposal is in the line of these works, which we detail in what follows.

Works on Mining Role Cardinality for Individuals. Several works consist in enriching the set of assertions on individuals (ABox), and we can distinguish the *endogenous* approaches [9] relying on the assertions already present in the ABox, from the *exogenous* approaches [13] relying on external sources. [9] shows that it is important to determine when a particular role (such as `parent`) is missing for a particular individual (such as *Obama*). Their proposal of Partial Completeness Assumption states that when at least one assertion about a role R is informed for an individual s, then all assertions for this role R are informed for this individual s. In [13], the authors benefit from text mining applied on Wikipedia for improving the completeness of individuals described in Wikidata. This exogenous approach relies on syntactical patterns to identify cardinalities on individuals. More generally, in [8], the authors propose various kinds of endogenous and exogenous heuristics for characterizing the completeness of individuals, called Completeness Oracles, as for instance taking into account the popularity of individuals (i.e., a famous individual is more likely to have complete information). Our proposal is endogenous as it processes the facts already contained in the KB that we want to enrich. Nevertheless, it does not characterize the role cardinality for a specific *individual* but for a *concept*. It is therefore more general as the constraints for concept C apply for all the individuals of C.

Works on Mining Role Cardinality for Concepts. Other proposals have focused on the enrichment of the schema part (TBox) with new assertions or axioms allowing to partially or completely specify the cardinality of a role. In particular, several works [16,19] address the automated discovery of contextual keys in RDF datasets as it was done in relational databases. They find axioms stating that individuals of a concept C must have only one tuple of values for a given tuple of roles. The same kind of cardinality information is induced by [12]. Indeed, the authors propose to discover roles that are mandatory for individuals of a concept C. For this purpose, they compare the density of the role R for individuals of the concept C with the densities of R for other concepts in the concept hierarchy. Our proposal focuses on mining the maximum cardinality for a role R in a context C (if it exists). But, contrary to the previous work, we can get information about cardinalities greater than 1 (e.g., 2 parents for a child). To the best of our knowledge, [15] is the only work explicitly dedicated to the detection of minimum/maximum cardinalities. This approach proceeds in two stages: removal of outliers and calculation of bounds. Unfortunately, KBs are so incomplete that the filtering of outliers is ineffective (e.g., there are more children with only one parent than children with 2 parents). Moreover, their filtering method implicitly assumes that the cardinalities follow a normal distribution, or a distribution

that is moderately asymmetric, which is not always the case (see the examples of Table 1). Consequently, for DBpedia their approach finds that a person has at most 2 years of birth (instead of 1) and 3 parents (instead of 2); and a football match has 3 teams (instead of 2). It is also important to note that the method extracts a cardinality constraint for every concept and role of the KB, whatever the number of observations and the distribution (e.g., a constraint for team is found in the context $\top$). Thus, many of these constraints are not significant. On the contrary, our approach benefits from Hoeffding's inequality for ensuring statisical significance. Finally, contrary to our approach, the authors do not envisage an algorithm to systematically explore the roles and concepts of the KB. An exploration strategy is yet crucial and not trivial in practice due to the huge search space.

Interest of Role Cardinality. Whatever the approach, all information extracted about role cardinalities is useful for improving many methods, as they reduce the uncertainty imposed by the open-world assumption. [9,20] show the necessity of reducing this uncertainty for data mining applied to KB. In particular, [8,9] propose to benefit from the previously mentioned Partial Completeness Assumption for improving the confidence estimation of association rules. More recently, [20] has further improved the confidence estimation of a rule by exploiting the bounds on the cardinality for an individual. Data mining is not the only field where insights about cardinalities are useful. [3,4,17] and more recently [10] propose to characterize query answers benefiting from the completeness degree of the queried data. Most of these methods can therefore directly benefit from the constraints that we investigate in this paper.

3 Preliminaries and Problem Formulation

3.1 Basic Notations

For talking about KB components, we use Description Logics (DL) [2] terminology. For instance DBpedia is a KB $\mathcal{K} = (\mathcal{T}, \mathcal{A})$, where $\mathcal{T}$ denotes its TBox and $\mathcal{A}$ denotes its ABox. One example of assertion in $\mathcal{T}$ is $\texttt{Artist} \sqsubseteq \texttt{Person}$, meaning that the concept `Artist` is subsumed by the concept `Person`, i.e. all artists are persons. $\mathcal{T}$ also includes assertions like $\exists\texttt{birthYear} \sqsubseteq \texttt{Person}$, meaning that the role `birthYear` is defined for persons. Note that the only part of the TBox used by our approach is the named hierarchies of concepts. Besides, $\texttt{Person}(\mathit{Obama})$ and $\texttt{birthYear}(\mathit{Obama}, 1961)$ are assertions of DBpedia's ABox $\mathcal{A}$. The former indicates that *Obama* is a person, while the latter states that *Obama* was born in 1961. In this paper, we assume that a KB $\mathcal{K}$ contains only one hierarchy of concepts and we use the general top concept $\top$ which subsumes every concept in $\mathcal{K}$. In DL, a maximum cardinality M on the role R may be represented using the numerical restriction constructor $\leq M\ R$. $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ implies[3] the constraint $\top \sqsubseteq (\leq M\ R)$, if for all subjects s, the number of objects o such that $R(s, o)$

[3] DL formal semantics are given in terms of interpretations, see [2].

belongs to $\mathcal{K}$ (i.e., $R(s,o) \in \mathcal{A}$ or $R(s,o)$ can be inferred from $\mathcal{T}$ and $\mathcal{A}$) is equal to or fewer than M.

We focus on cardinality constraints that are *contextual*, as stated in Definition 1. Intuitively, these constraints are not necessarily satisfied for all subjects of a role R, but for all the subjects of R that belong to a concept C.

Definition 1 (Contextual Constraint). *Given an integer $M \geq 1$, a role R and a concept C of a KB $\mathcal{K}$, a contextual maximum cardinality constraint defined on R for C is an expression of the form: $C \sqsubseteq (\leq M\ R)$.*

The concept C is called the context of the constraint $C \sqsubseteq (\leq M\ R)$. For example, the contextual constraint $\texttt{Person} \sqsubseteq (\leq 1\ \texttt{birthYear})$ means that each person has at most 1 birth year, while $\texttt{FootballMatch} \sqsubseteq (\leq 2\ \texttt{team})$ means that a football match has at most 2 teams. Note that asserting that an artist has at most one year of birth (i.e., $\texttt{Artist} \sqsubseteq (\leq 1\ \texttt{birthYear})$) is true, but less general than $\texttt{Person} \sqsubseteq (\leq 1\ \texttt{birthYear})$ because $\texttt{Artist} \sqsubset \texttt{Person}$. Similarly, asserting that 1,000 is a maximum cardinality for the parent role (i.e., $\texttt{Person} \sqsubseteq (\leq 1,000\ \texttt{parent})$) is true, but less specific than $\texttt{Person} \sqsubseteq (\leq 2\ \texttt{parent})$. We want to discover contextual maximum cardinality constraints that have a context as general as possible and a cardinality as small as possible. For this purpose, we introduce the notion of minimal contextual constraint:

Definition 2 (Minimal Contextual Constraint). *The contextual constraint $\gamma_1 : C_1 \sqsubseteq (\leq M_1\ R)$ is more general than the contextual constraint $\gamma_2 : C_2 \sqsubseteq (\leq M_2\ R)$, denoted by $\gamma_2 \sqsubset \gamma_1$, iff $C_2 \sqsubset C_1$[4] and $M_1 \leq M_2$, or $C_2 \equiv C_1$ and $M_1 < M_2$. For a given set of contextual constraints Γ, constraint $\gamma_1 \in \Gamma$ is minimal in Γ if there is no constraint $\gamma_2 \in \Gamma$ more general than γ_1: $(\nexists \gamma_2 \in \Gamma)(\gamma_1 \sqsubset \gamma_2)$.*

The notion of minimality restricts the mining to a set of constraints that is not redundant, meaning that we do not want to extract a maximum cardinality constraint γ_2 if it is logically implied by another maximum cardinality constraint γ_1. More precisely, it is easy to see that if a maximum cardinality constraint $\gamma_1 : C_1 \sqsubseteq (\leq M_1\ R)$ is more general than a maximum cardinality constraint $\gamma_2 : C_2 \sqsubseteq (\leq M_2\ R)$, then for all interpretation $\mathcal{I}$ of a KB $\mathcal{K}$, if $\mathcal{I}$ is a model of γ_1, then $\mathcal{I}$ is also a model of γ_2. Indeed, if $\mathcal{I}$ is a model of γ_1, we have $C_1^{\mathcal{I}} \subseteq \{o : \#\{o' : (o,o') \in R^{\mathcal{I}}\} \leq M_1\}$. Moreover, since γ_1 is more general than γ_2, we have $C_2^{\mathcal{I}} \subseteq C_1^{\mathcal{I}}$ and $M_1 \leq M_2$. Thus, we have $C_2^{\mathcal{I}} \subseteq C_1^{\mathcal{I}} \subseteq \{o : \#\{o' : (o,o') \in R^{\mathcal{I}}\} \leq M_1\} \subseteq \{o : \#\{o' : (o,o') \in R^{\mathcal{I}}\} \leq M_2\}$, which shows that $\mathcal{I}$ is a model of γ_2.

Note that our method relies on a named concept hierarchy for exploring possible contexts and using their subsumption relations. However, it is possible to generate such a hierarchy to explore more complex contexts in a pre-processing step. Such an approach is useful to analyze data by expressing the background knowledge of an expert through an analytical hierarchy.

[4] We denote $C \sqsubset C'$ when $C \sqsubseteq C'$ and $C' \not\sqsubseteq C$.

3.2 Problem Statement

Considering the statistics in DBpedia provided by Table 1, we do not want to discover the contextual constraints $\texttt{Person} \sqsubseteq (\leq 6\ \texttt{birthYear})$ or $\texttt{Person} \sqsubseteq (\leq 5\ \texttt{parent})$ even if these constraints are satisfied and minimal in $\mathcal{K}$. We would intend to extract the contextual constraints $\texttt{Person} \sqsubseteq (\leq 1\ \texttt{birthYear})$ or $\texttt{Person} \sqsubseteq (\leq 2\ \texttt{parent})$. Therefore, as defined in [14], we assume an ideal description of the world or ideal KB, denoted $\mathcal{K}^*$, in the sense that $\mathcal{K}^*$ is *correct* (it does not contain any inconsistancies) and *complete*. Note that in general, we have neither $\mathcal{K} \subseteq \mathcal{K}^*$, nor $\mathcal{K}^* \subseteq \mathcal{K}$, because $\mathcal{K}$ is inconsistent or incomplete. In this context, our problem can be formalized as follows:

Problem 1. Given a knowledge base $\mathcal{K}$, we aim at discovering the set of all contextual maximum cardinality constraints $C \sqsubseteq (\leq M\ R)$ where C and R are concept and role of $\mathcal{K}$, that are *satisfied* in $\mathcal{K}^*$ and *minimal* with respect to the concept hierarchy of $\mathcal{K}$.

In order to solve Problem 1 we have to deal with the two following challenges: (i) discover constraints that would be satisfied in $\mathcal{K}^*$ whereas this knowledge base is hypothetical and unknown (see Sect. 4), and (ii) efficiently explore the search space knowing that the number of possible contextual maximum cardinality constraints is huge (see Sect. 5).

4 Detecting Significant Maximum Cardinalities

This section use a probability framework relying on the hypothesis that the degree of completeness of a role is in general higher than its level of inconsistencies. For instance, this assumption is reasonable for DBpedia. Indeed, even if it is difficult to evaluate the completeness and the semantic accuracy of a knowledge base because it requires a gold standard [5], several results of the literature tend to show that the semantic accuracy of DBpedia is better than its completeness [7].

More formally, let us assume that M is the *true* maximum cardinality of the role R in the context C, meaning that the maximum cardinality constraint $\gamma : C \sqsubseteq (\leq M\ R)$ is satisfied in $\mathcal{K}^*$. In practice, the ideal KB $\mathcal{K}^*$ is unknown and we only have a sample $\mathcal{K}$ of the reality. Let X be the random variable that denotes for a subject s the number of assertions $R(s, o)$ observed in $\mathcal{K}$. We assume that:

- The level of inconsistencies in $\mathcal{K}$ is not significant, i.e. the probability $\mathbf{P}(X > M)$ to observe a cardinality greater than M for role R is low. For example, in Table 1, we can see that 85 individuals of context `Person` have more than 2 parents, but they represent less than 0.43% of the observed individuals.
- The degree of completeness (present roles) is significantly higher, i.e. the probability $\mathbf{P}(X = M)$ to observe the maximum cardinality M is significantly higher than $\mathbf{P}(X > M)$. For example, in Table 1, we can see that 9,342 individuals of context `Person` have 2 parents, which represents more than 46.7% of the observed individuals.

Under these hypotheses, the following property states that if M is the true maximum cardinality of the role R in the context C, then M is the integer i that maximizes the conditional probability $\mathbf{P}(X = i|X \geq i)$:

Property 1. Let M be the *true* maximum cardinality of the role R in the context C. If $\mathbf{P}(X = M) \geq \lambda$ and $\mathbf{P}(X > M) \leq \epsilon$, then we have $\mathbf{P}(X = M|X \geq M) \geq \frac{\lambda}{\lambda+\epsilon}$ and $\mathbf{P}(X = i|X \geq i) \leq (1 - \lambda)$ for $i \in [1..M[$. Moreover, if $\lambda > 1/2(\sqrt{\epsilon^2 + 4\epsilon} - \epsilon)$, we have: $M = \arg\max_{i \in \mathbb{N}^+} \{\mathbf{P}(X = i|X \geq i) : \mathbf{P}(X = i) > \epsilon\}$.

Due to lack of space, we omit the proofs. Assuming an inconsistency level ϵ equal to 0.1% (resp. 1%), Property 1 states that it is possible to detect a true maximum cardinality if the degree of completeness λ is greater than $1/2(\sqrt{0.001^2 + 4 \cdot 0.001} - 0.001) = 3.2\%$ (resp. 9.5%). Moreover, a true maximum cardinality constraint M will be detected if $\mathbf{P}(X = M|X \geq M) \geq \frac{\lambda}{\lambda+\epsilon} \geq 97\%$ (resp. 90%). Finally, note that when there is no inconsistency (i.e., $\mathbf{P}(X > M) = 0$ and $\epsilon = 0$), if M is a true maximum cardinality, then $\mathbf{P}(X = M|X \geq M) = 1$.

Now, based on this assumption, we define in Sect. 4.1 the measure of *likelihood* to detect maximum cardinality constraints, and show how to use Hoeffding's inequality to obtain more accurate decisions. Besides, we introduce in Sect. 4.2 the notion of *significant constraint*.

4.1 Likelihood Measure

We now introduce the notion of likelihood to measure a frequency estimation of the conditional probability $\mathbf{P}(X = i|X \geq i)$ involved in Property 1 (for deciding whether a cardinality i for the role R in the context C is likely to be maximum):

Definition 3 (Likelihood). *Given a knowledge base $\mathcal{K}$, the likelihood of the maximum cardinality i of the role R for the context C is the ratio defined as follows: $\tau_i^{C,R}(\mathcal{K}) = \frac{n_i^{C,R}}{n_{\geq i}^{C,R}}$ if $n_{\geq i}^{C,R} > 0$ (0 otherwise) where $n_i^{C,R}$ (resp. $n_{\geq i}^{C,R}$) is the number of individuals s of the context C such that i facts $R(s, o)$ (resp. i facts or more) are stated in $\mathcal{K}$.*

When the context and the role are clear, we omit them in notations. In that case, n_i, $n_{\geq i}$ and $\tau_i(\mathcal{K})$ respectively denote $n_i^{C,R}$, $n_{\geq i}^{C,R}$ and $\tau_i^{C,R}(\mathcal{K})$.

For example, let us consider the context `Person` and the role `parent`. Using Table 1, it is easy to see that $n_{\geq 2}^{\texttt{Person,parent}} = 9,477$ $(9,477 = 9,392 + 75 + 9 + 1)$. Thereby, the likelihood $\tau_2^{\texttt{Person,parent}}(\mathcal{K})$ is 0.991 (i.e., $9,392/9,477$). Note that this measure ignores the $10,643$ persons that have only one informed parent (to evaluate if 2 is the true maximum cardinality for parents). Then, it is also easy to see that we have $\tau_6^{\texttt{Person,parent}}(\mathcal{K}) = 1$, whereas 6 is not the true maximum cardinality for the role `parent`. Intuitively, if the likelihood $\tau_6^{\texttt{Person,parent}}(\mathcal{K}) = 1$ does not make sense, it is due to an insufficient number of individuals for reinforcing this hypothesis (here, only 1 individual has 6 parents). In general, the estimation of $\mathbf{P}(X = i|X \geq i)$ by $\tau_i(\mathcal{K})$ must be corrected to be statistically

valid. For this purpose, we benefit from the Hoeffding's inequality [11] which has the advantage of being true for any distribution. It provides an upper bound on the probability that an empirical mean (in our case, a likelihood $\tau_i(\mathcal{K})$) deviates from its expected value (the conditional probability $\mathbf{P}(X = i|X \geq i)$) by more than a given amount. More formally, we have the following property:

Property 2 (Lower bound). Given a knowledge base $\mathcal{K}$ and a confidence level $1 - \delta$, assuming that all the observations are independently and identically distributed, the conditional probability $\theta_i = \mathbf{P}(X = i|X \geq i)$ is greater than the pessimistic likelihood $\widetilde{\tau}_i(\mathcal{K})$ defined by (if $n_{\geq i} > 0$):

$$\widetilde{\tau}_i(\mathcal{K}) = \max\left\{\frac{n_i}{n_{\geq i}} - \sqrt{\frac{\log(1/\delta)}{2n_{\geq i}}}, 0\right\}$$

with a probability greater than $(1 - \delta)$, i.e. $\mathbf{P}(\theta_i \geq \widetilde{\tau}_i(\mathcal{K})) \geq (1 - \delta)$.

This property provides us an efficient tool to make confident decisions. For instance, for the role `parent` in Table 1, we observe that the correction strongly reduces the likelihood $\tau_i(\mathcal{K})$ for cardinalities 3, 4 and 6 (e.g., $\widetilde{\tau}_6^{\texttt{Person,parent}}(\mathcal{K}) = 0$). Conversely, we have $\widetilde{\tau}_2^{\texttt{Person,parent}}(\mathcal{K}) = 0.975$, a strong indicator to consider that 2 is the true maximum cardinality for the role `parent` in the context `Person`.

4.2 Significant Maximum Cardinality

Using Properties 1 and 2, we finally propose to detect a maximum cardinality M for a confidence level $1 - \delta$ if (i) the pessimistic likelihood $\widetilde{\tau}_M(\mathcal{K})$ is maximum, i.e. $\widetilde{\tau}_M(\mathcal{K}) = \max_{i>0} \widetilde{\tau}_M(\mathcal{K})$, and (ii) the pessimistic likelihood $\widetilde{\tau}_M(\mathcal{K})$ is greater than a minimum likelihood threshold min_τ. Based on this heuristic, we introduce the notion of *significant* maximum cardinality constraint:

Definition 4 (Significant Constraint). *Given a minimum likelihood threshold min_τ, a confidence level $1-\delta$ and a knowledge base $\mathcal{K}$, a contextual maximum cardinality constraint $C \sqsubseteq (\leq M\ R)$ is significant w.r.t. $\mathcal{K}$ iff $\widetilde{\tau}_M(\mathcal{K}) \geq min_\tau$ and $\widetilde{\tau}_M(\mathcal{K}) = max_{i\geq 1}\widetilde{\tau}_i(\mathcal{K})$.*

Compared to Property 1, note that in our heuristic, we do not test whether $\widetilde{\tau}_M$ is greater than ϵ, or not. However, it is easy to see that if $\widetilde{\tau}_M = \tau_M - \sqrt{\frac{\log(1/\delta)}{2n_{\geq M}}} \geq min_\tau$, then we necessarily have $n_{\geq M} \geq \frac{\log(1/\delta)}{2(1-min_\tau)^2}$, which guarantees that we will not make a decision if the number of observations $n_{\geq M}$ is too low. For example, with $1 - \delta = 99\%$ and $min_\tau = 0.97$, we will consider that M is a true maximum cardinality only if $n_{\geq M} \geq 2,558$.

In DBpedia for a confidence level $1 - \delta = 99\%$ and a threshold $min_\tau = 0.97$, we observe that the detected maximum cardinalities of the roles `birthYear` and `parent` in the context `Person` are 1 and 2 respectively (bold values in Table 1). Interestingly, with these same thresholds, no maximum cardinality is detected for the role `team` when no context is considered. This is because this role is used

both to inform the teams to which a player has belonged and the teams present in a sport event. Thence, our method manages to detect the cardinality of 2 in the context of football matches.

By Definition 4, if a constraint is *significant* w.r.t. $\mathcal{K}$, it means that its pessimistic likelihood is greater than min_τ and that it is probably satisfied in $\mathcal{K}^*$ (using Properties 1 and 2). Now, our problem is expressed as follows:

Problem 2. Given a knowledge base $\mathcal{K}$ satisfying the assumptions expressed in Sect. 4 about its consistency and its completeness, a confidence level $1-\delta$ and a minimum likelihood threshold min_τ, we aim at discovering the set of all contextual maximum cardinality constraints $C \sqsubseteq (\leq M\ R)$ where C and R are concept and role of $\mathcal{K}$, that are *significant* w.r.t. $\mathcal{K}$ and *minimal* w.r.t. the concept hierarchy defined in the TBox of $\mathcal{K}$.

5 Extracting Maximum Cardinality Constraints

5.1 Pruning Criteria

For discovering all the contextual constraints of a knowledge base $\mathcal{K}$, a naive approach would consist in testing each role for each concept with our detection method. If N_C is the number of concepts and N_R the number of roles, this naive approach would require $N_C \times N_R$ tests. This is unfeasible for large knowledge bases such as DBpedia, containing more than 483k concepts and 60k roles. We design two pruning criteria (Properties 3 and 4) taking advantage of the two conditions that a constraint γ must satisfy to be mined: (i) the constraint γ has to be *significant* i.e., its pessimistic likelihood has to be greater than the minimum likelihood threshold min_τ, and (ii) the constraint γ has to be *minimal* with respect to the hierarchy of concepts defined in the TBox of $\mathcal{K}$.

First, we show that a constraint $C \sqsubseteq (\leq M\ R)$ cannot be significant if the number of individuals of the context C in $\mathcal{K}$ is too small. Indeed, if $|C|$ is too small, the confidence interval computed with Hoeffding's inequality is very large and consequently, the pessimistic likelihood is lower than the minimum threshold min_τ. This intuition is formally presented in this property:

Property 3 (Significance pruning). Given a confidence level $1-\delta$ and a minimum likelihood threshold min_τ, if one has $|C \sqcap (\exists R.\top)| < \frac{\log(1/\delta)}{2(1-min_\tau)^2}$ for the context C and the role R, then no contextual constraint $C' \sqsubseteq (\leq M\ R)$ with $C' \sqsubseteq C$ can be significant w.r.t. the knowledge base $\mathcal{K}$.

This property is very important to reduce the search space because if the number of individuals in $\mathcal{A}$ that belong to $C \sqcap (\exists R.\top)$, for a context C and a role R, is not large enough (if it is lower than $\log(1/\delta)/2(1-min_\tau)^2$), then it is impossible to find a significant constraint $C' \sqsubseteq (\leq M\ R)$ where C' is a concept more specific than C in the hierarchy of $\mathcal{K}$. For example, we use a minimum likelihood threshold min_τ of 97% and a confidence $1-\delta$ of 99% to extract constraints in DBpedia (see experimental sections), which means that at least

2,558 observations are needed for a role R in a context C. For this reason, since there are only 896 facts for the role `beatifiedDate` describing the context `Person`, we are sure that it is not necessary to explore this role for the sub-concepts like `Artist` or `Scientist`.

Assume now that we have extracted the constraint $C \sqsubseteq (\leq 1\ R)$ from the knowledge base $\mathcal{K}$. It is not possible to find another *minimal* constraint $C' \sqsubseteq (\leq M'\ R)$ with a context C' more specific than C because the cardinality M' cannot be smaller than 1. This property, which is a direct consequence of minimality (see Definition 2), is formalized as follows:

Property 4 (Minimality pruning). Let Γ be a set of contextual maximum cardinality constraints. If Γ contains a contextual constraint $C \sqsubseteq (\leq 1\ R)$, then no contextual constraint $C' \sqsubseteq (\leq M'\ R)$ with $C' \sqsubset C$ can be minimal in Γ.

Property 4 is also useful to reduce the search space because if a constraint $C \sqsubseteq (\leq 1\ R)$ has been detected as significant, then it is useless to explore all the constraints $C' \sqsubseteq (\leq M'\ R)$ where $C' \sqsubset C$. As soon as the constraint `Person` $\sqsubseteq (\leq 1$ `birthYear`$)$ has been detected (meaning than a person has at most one birth year), it is no longer necessary to explore the constraint `Artist` $\sqsubseteq (\leq M$ `birthYear`$)$ which is more specific.

5.2 C3M: Contextual Cardinality Constraint Mining

Properties 3 and 4 are implemented in our algorithm called C3M (*C3M* for *Contextual Cardinality Constraint Mining*). Its main function, called *C3M-Main*, takes as input a knowledge base $\mathcal{K}$, a confidence level $1 - \delta$ and a minimum likelihood threshold min_τ. The exploration of the search space is performed independently for each role R of the knowledge base $\mathcal{K}$ (see the main loop of Algorithm 1 at line 2). In a first phase, given a role R of $\mathcal{K}$, Algorithm 1 carries out a depth-first exploration of cardinality constraints for R (line 4). This exploration starts from the top concept of $\mathcal{K}$, denoted by $\top$, by calling the recursive function *C3M-Explore*. Because the concepts of $\mathcal{K}$ may have multiple more general concepts, the set Γ_R of maximum cardinality constraints returned by function *C3M-Explore* may contain constraints that are not minimal. Therefore, in a second phase (line 6), the function *C3M-Main* checks for each constraint $\gamma \in \Gamma_R$ if Γ_R contains a constraint γ' that is more general than γ. When it is not the case constraint γ is added to the set of maximum cardinality constraints Γ_m that are minimal. Γ_m is finally returned by function *C3M-Main* (line 8).

The recursive function *C3M-Explore* benefits from the pruning criteria presented in Properties 3 and 4 during a depth-first exploration of the search space. First, it evaluates if the number of observations in $C \sqcap (\exists R.\top)$ is sufficiently important. If it is not the case, we know that there is no maximum cardinality constraint $C' \sqsubseteq (\leq M\ R)$ with $C' \sqsubseteq C$ that can be significant w.r.t. $\mathcal{K}$ (see Property 3) and the depth-first exploration is stopped (line 2 of Algorithm 2). Otherwise, the pessimistic likelihood $\widetilde{\tau}_i$ is computed for each cardinality value i

Algorithm 1. C3M-Main

Input: A knowledge base $\mathcal{K}$, a confidence level $1-\delta$ and a minimum likelihood threshold min_τ

Output: The set Γ_m of all maximum cardinality constraints that are significant and minimal w.r.t. $\mathcal{K}$

1: $\Gamma_m := \emptyset$
2: **for all** role in $\mathcal{K}$ **do**
3: {Depth-first exploration of maximum cardinality constraints}
4: $\Gamma_R :=$ *C3M-Explore*$(\mathcal{K}, R, \top, \infty, \delta, min_\tau)$
5: {Computation of maximum cardinality constraints that are minimal}
6: $\Gamma_m := \{\gamma \in \Gamma_R : (\nexists \gamma' \in \Gamma_R)(\gamma \sqsubset \gamma')\} \cup \Gamma_m$
7: **end for**
8: **return** Γ_m

(lines 4–6) and the most likely cardinality i_M is selected (line 7). If the corresponding pessimistic likelihood $\widetilde{\tau}_{i_M}$ is lower than min_τ, it means that no maximum cardinality constraint is detected (for this level of the hierarchy of $\mathcal{K}$) and i_M is set to ∞ (line 8). Otherwise, if i_M is strictly lower than M (the maximum cardinality detected at a previous level of the hierarchy), it means that we detect a maximum constraint cardinality $\gamma : C \sqsubseteq (\leq i_M\ R)$ that is *potentially* minimal. As already mentioned, as a concept of the knowledge base $\mathcal{K}$ may have multiple super-concepts, we will have to check whether γ is really minimal in the second phase of function *C3M-Main*. Finally, using Property 4, we know that if $i_M = 1$, it is not necessary to explore the descendants $C' \sqsubset C$ to detect other constraints $C' \sqsubseteq (\leq M'\ R)$. Otherwise, *C3M-Explore* is recursively called (line 12) to explore all the direct sub-concepts of C (identified using the hierarchy in the TBox of $\mathcal{K}$).

Theorem 1. *Given a knowledge base $\mathcal{K}$, a confidence level $1-\delta$ and a minimum likelihood min_τ, our algorithm C3M-Main returns the set of all contextual cardinality constraints $C \sqsubseteq (\leq M\ R)$ that are significant w.r.t. $\mathcal{K}$ and minimal w.r.t. the hierarchy of concepts defined in the TBox of $\mathcal{K}$.*

Theorem 1 straightforwardly stems from Properties 3 and 4. Although these pruning criteria are not heuristic, we will see in the experimental section that algorithm *C3M-Main* is efficient enough to handle knowledge bases as large as DBpedia. Note that we have implemented the functions *C3M-Main* and *C3M-Explore* (client side) such that they consume a SPARQL endpoint (server side) to query the knowledge base $\mathcal{K}$. More precisely, given a context C and a role R, a SPARQL query is built and executed to compute the cardinality distribution $n_i^{C,R}$ $(i \in \mathbb{N})$, which is useful for calculating pessimistic likelihoods (see line 5 of Algorithm 2). Therefore, for each role R in $\mathcal{K}$, the server side executes N_C queries where N_C represents the number of concepts in the hierarchy of concepts of $\mathcal{K}$. It means that the complexity of our approach in number of queries is in $\mathcal{O}(N_C)$. On the other hand, on the client side (where the functions *C3M-Main* and *C3M-Explore* are executed), given a role R of $\mathcal{K}$, the complexity of our approach (in

Algorithm 2. C3M-Explore

Input: A knowledge base $\mathcal{K}$, a role R, a context C, a cardinality M, a confidence level $1-\delta$ and a minimum likelihood threshold min_τ

Output: A set Γ of constraints

$\alpha := \frac{log(1/\delta)}{2(1-min_\tau)^2}$ and $n_{\geq 0}^{C,R} := |C \sqcap (\exists R.\top)|$
if $(n_{\geq 0}^{C,R} < \alpha)$ **then return** $\emptyset$
$\Gamma := \emptyset$ and $i_{max} := \arg\max_{i \in \mathbb{N}} \{n_i^{C,R} > 0\}$
for all $i \in [1..min\{M, i_{max}\}]$ **do**
    $\widetilde{\tau}_i := \max \left\{ \frac{n_i^{C,R}}{n_{\geq i}^{C,R}} - \sqrt{\frac{\log(1/\delta)}{2n_{\geq i}^{C,R}}}; 0 \right\}$
end for
$i_M := \arg\max_{i \in [1..min\{M, i_{max}\}]} \{\widetilde{\tau}_i\}$
if $(\widetilde{\tau}_{i_M} < min_\tau)$ **then** $i_M := \infty$
if $(i_M < M)$ **then** $\Gamma := \{C \sqsubseteq (\leq i_M\ R)\}$
if $(i_M > 1)$ **then**
    for all direct sub-concept $C' \sqsubset C$ not yet explored **do**
        $\Gamma := \Gamma \cup$ *C3M-Explore*$(\mathcal{K}, R, C', i_M, \delta, min_\tau)$
    end for
end if
return Γ

number of operations) is in $\mathcal{O}(N_C \times i_{max})$ where $i_{max} = \arg\max_{i \in \mathbb{N}} \{n_i^{\top,R} > 0\}$. Intuitively, i_{max} represents the maximum integer for which there is at least one subject s such that i_{max} facts $R(s, o)$ belong to $\mathcal{K}$.

6 Experiments

The goal of this experimental study is mainly to evaluate the scaling of the C3M algorithm with a large knowledge base, the interest of minimality and the precision of the mined constraints. In this paper, we present and analyze experimental results using DBpedia, which contains more than 500 million triples with more than 480k distinct concepts and 60k distinct roles. The Github repository of C3M (see footnotes) also provides results obtained from 3 other SPARQL endpoints: YAGO, BNF and EUROPEANA.

Our algorithm is implemented in Java with the Apache Jena Library, and directly queries the KB via its SPARQL endpoint[5]. Note that we virtually add an element $\top$ that subsumes all concepts without parents including `owl:Thing`, and the confidence level is $1 - \delta = 99\%$ for all experiments[6]. Figure 1 varies the minimum likelihood threshold min_τ from 0.90 to 0.99 to observe the evolution of the collection of contextual maximum cardinality constraints.

[5] http://jena.apache.org and https://dbpedia.org.

[6] The results for $min_\tau = 0.97$ and the ground truth used to evaluate the precision are available at https://github.com/asoulet/c3m.

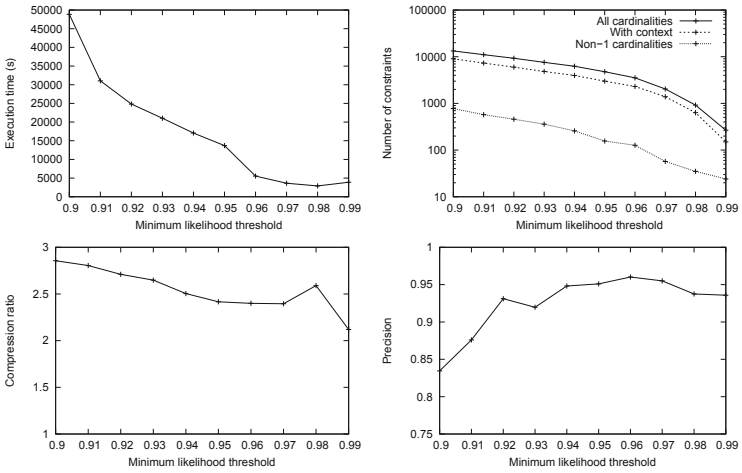

Fig. 1. Impact of the minimum likelihood threshold

Scalability. Figure 1 (left top) reports the execution time, which increases very rapidly when the likelihood threshold decreases. This is due to a very rapid increase of the size of the search space because the pruning properties are less selective. As a result, the number of extracted contextual constraints also increases with the decrease of the threshold min_τ as shown in Fig. 1 (right top). More precisely, it reports the total number of mined constraints, the number of constraints with a non-$\top$ context (i.e., with context different from $\top$), and the number of non-1 constraints (i.e., with maximum cardinality greater than 1). First, it is clear that a majority of constraints have 1 as cardinality. For a minimum likelihood threshold equal to 0.97, there are 1,979 constraints with 1 as maximum cardinality (see Fig. 2 (left) that details the distribution of constraints with cardinality). Second, we also observe that most of constraints have a non-$\top$ context that shows the usefulness of our approach based on contexts. For $min_\tau = 0.97$, Fig. 2 (right) plots the distribution of the constraints with the level of their context in the DBpedia hierarchy.

Minimality. Figure 1 (left bottom) plots the compression ratio due to minimality (i.e., number of minimal and non-minimal constraints divided by the number of minimal constraints) by varying the likelihood threshold. Interestingly, the reduction of the number of constraints thanks to minimality is important regardless of the threshold (between 2 and 3 times smaller). It is slightly less effective when the likelihood threshold is high, but much fewer constraints are identified. As a reminder, the non-minimal pruned constraints are not informative because redundant with more general ones. In other words, they are not useful for an inference system and in addition, they reduce the readability of the extraction for end users.

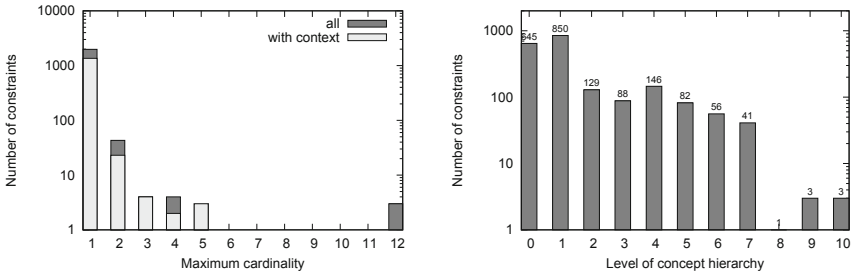

Fig. 2. Distribution of constraints for $min_\tau = 0.97$

Precision. In order to evaluate the quality of the mined constraints, we built a ground truth from a set $\mathcal{C}^*$ of 5,041 constraints selected from the 13,313 constraints extracted with $min_\tau = 0.90$. We first used common sense knowledge and information from the DBpedia pages to determine the maximum cardinalities of certain relations. For instance, since we have a single birth, the maximum cardinality for all birth dates and places has been set to 1. For some relations like `rdfs:label` or `rdfs:abstract`, the maximum cardinality has been set to 12 according to the documentation[7]. In a second step, we automatically extended the maximum cardinality constraints to the different contexts. The set $\mathcal{C}^*$ covers 667 distinct roles and 2,150 distinct concepts. Thereby, the precision of a set of constraints $\mathcal{C}$ corresponds to the proportion of correct constraints out of the number of constraints that are annotated (i.e., $\mathcal{C} \cap \mathcal{C}^*$). Figure 1 (right bottom) plots the precision of the set of constraints returned by C3M according to the minimum likelihood threshold min_τ[8]. We observe that precision increases with this threshold, but drops off for thresholds greater than 0.96. This is due to correct cardinality constraints which are not recognized as the needed number of individuals is too high. However, it is important to note that this decrease is not very significant because the number of mined constraints becomes very small for thresholds greater than 0.96. Interestingly, for a threshold greater than or equal to 0.94, the precision of our approach is excellent since about 95% of the constraints are correct.

We also qualitatively analyzed the maximum cardinality constraints for a minimum likelihood threshold equal to 0.97. We observe that the erroneous constraints often result from construction or representation biases. For instance, the method found the constraint `http://schema.org/School` $\sqsubseteq$ ($\leq$2 `country`) that is wrong because a school is located in a single country. But we observe in DBpedia that many English schools are attached to both England and the United Kingdom. It is clear that a single affiliation to England (part of the United Kingdom) would have been sufficient. Besides, at physical level, while each individual has

[7] https://wiki.dbpedia.org/services-resources/datasets/dbpedia-datasets.

[8] We do not compare our method with [15] because in the case of DBpedia, this method systematically returns a *wrong maximum* cardinality for all constraints.

a unique date of birth, we identify a cardinality of 2 because many dates are represented with two distinct encoding formats.

To summarize, our approach scales well on DBpedia with about 500 million triples thanks to the advanced pruning techniques used by C3M. The majority of the extracted constraints have a context demonstrating the interest of benefiting from the concept hierarchy of the knowledge base. Importantly, the precision of the mined constraints is about 95% for $min_\tau \geq 0.94$.

7 Conclusion

This paper provides the first proposal for a complete exploration of significant constraints of maximum cardinality in a knowledge base. We show how to find, from a knowledge base $\mathcal{K}$ that satisfies assumptions about its completeness and consistence degrees, a minimal set of contextual constraints $C \sqsubseteq (\leq M\ R)$ that are *significant*, i.e. that can be expected to occur in reality. Our experiments demonstrate the feasibility of a systematic exploration of large knowledge bases such as DBpedia (about 500 million triples) for the discovery of minimal contextual constraints of maximum cardinality thanks to the C3M algorithm. With a high minimum likelihood threshold, the precision of the mined constraints is about 95%, which is excellent. Additionally, the minimality exploited by our algorithm drastically reduce the number of obtained constraints, so that they can be manually analyzed by end users. In future work, we would intend to extend our approach to minimum cardinality constraints. This task is not completely symmetrical because under the open-world assumption, it is difficult to know if facts are missing or if the minimum cardinality is reached. For instance, a majority of people have only one informed parent in DBpedia but, of course, the true minimum cardinality is 2. Another future work is to improve C3M by benefiting more from reasoning capabilities. For the moment, we take into account the hierarchy of concepts to reduce the set of constraints, but we could improve our approach by fully exploiting OWL (e.g., with equivalent classes or properties).

Acknowledgements. This work was partially supported by the grant ANR-18-CE38-0009 ("SESAME").

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52
2. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, New York (2003)

3. Darari, F., Nutt, W., Pirrò, G., Razniewski, S.: Completeness statements about RDF data sources and their use for query answering. In: Alani, H., et al. (eds.) ISWC 2013. LNCS, vol. 8218, pp. 66–83. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41335-3_5
4. Darari, F., Razniewski, S., Prasojo, R.E., Nutt, W.: Enabling fine-grained RDF data completeness assessment. In: Bozzon, A., Cudre-Maroux, P., Pautasso, C. (eds.) ICWE 2016. LNCS, vol. 9671, pp. 170–187. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-38791-8_10
5. Debattista, J., Lange, C., Auer, S., Cortis, D.: Evaluating the quality of the LOD cloud: an empirical investigation. Semant. Web **9**(6), 859–901 (2018)
6. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečić, D.: Introducing wikidata to the linked data web. In: Mika, P., et al. (eds.) ISWC 2014. LNCS, vol. 8796, pp. 50–65. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11964-9_4
7. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Semant. Web **9**(1), 77–129 (2018)
8. Galárraga, L., Razniewski, S., Amarilli, A., Suchanek, F.M.: Predicting completeness in knowledge bases. In: Proceedings of the 10th ACM International Conference on Web Search and Data Mining, pp. 375–383. ACM (2017)
9. Galárraga, L.A., Teflioudi, C., Hose, K., Suchanek, F.: AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In: Proceedings of World Wide Web Conference, pp. 413–422. ACM (2013)
10. Galárraga, L., Hose, K., Razniewski, S.: Enabling completeness-aware querying in SPARQL. In: Proceedings of the 21st Workshop on the Web and Databases, pp. 19–22. ACM (2017)
11. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**(310), 13–20 (1963)
12. Lajus, J., Suchanek, F.M.: Are all people married? Determining obligatory attributes in knowledge bases. In: Proceedings of World Wide Web Conference, pp. 1115–1124 (2018)
13. Mirza, P., Razniewski, S., Darari, F., Weikum, G.: Enriching knowledge bases with counting quantifiers. In: Vrandečić, D., et al. (eds.) ISWC 2018. LNCS, vol. 11136, pp. 179–197. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00671-6_11
14. Motro, A.: Integrity = validity + completeness. ACM Trans. Database Syst. **14**(4), 480–502 (1989)
15. Muñoz, E., Nickles, M.: Mining cardinalities from knowledge bases. In: Benslimane, D., Damiani, E., Grosky, W.I., Hameurlain, A., Sheth, A., Wagner, R.R. (eds.) DEXA 2017. LNCS, vol. 10438, pp. 447–462. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64468-4_34
16. Pernelle, N., Saïs, F., Symeonidou, D.: An automatic key discovery approach for data linking. Web Semant.: Sci. Serv. Agents World Wide Web **23**, 16–30 (2013)
17. Razniewski, S., Korn, F., Nutt, W., Srivastava, D.: Identifying the extent of completeness of query answers over partially complete databases. In: Proceedings of the ACM SIGMOD, pp. 561–576. ACM (2015)
18. Soulet, A., Giacometti, A., Markhoff, B., Suchanek, F.M.: Representativeness of knowledge bases with the generalized Benford's law. In: Vrandečić, D., et al. (eds.) ISWC 2018. LNCS, vol. 11136, pp. 374–390. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00671-6_22

19. Symeonidou, D., Galárraga, L., Pernelle, N., Saïs, F., Suchanek, F.: VICKEY: mining conditional keys on knowledge bases. In: d'Amato, C., et al. (eds.) ISWC 2017. LNCS, vol. 10587, pp. 661–677. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68288-4_39
20. Pellissier Tanon, T., Stepanova, D., Razniewski, S., Mirza, P., Weikum, G.: Completeness-aware rule learning from knowledge graphs. In: d'Amato, C., et al. (eds.) ISWC 2017. LNCS, vol. 10587, pp. 507–525. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68288-4_30
21. Weikum, G., Hoffart, J., Suchanek, F.M.: Ten years of knowledge harvesting: lessons and challenges. IEEE Data Eng. Bull. **39**(3), 41–50 (2016)