Auckland 2019 ISWC

Mining Significant Maximum Cardinalities in Knowledge Bases

Arnaud Giacometti - Béatrice Markhoff - Arnaud Soulet

LIFAT – Université de Tours (France)

LIFAT EA 6300 LABORATOIRE D'INFORMATIQUE FONDAMENTALE ET APPLIQUÉE DE TOURS

université de TOURS

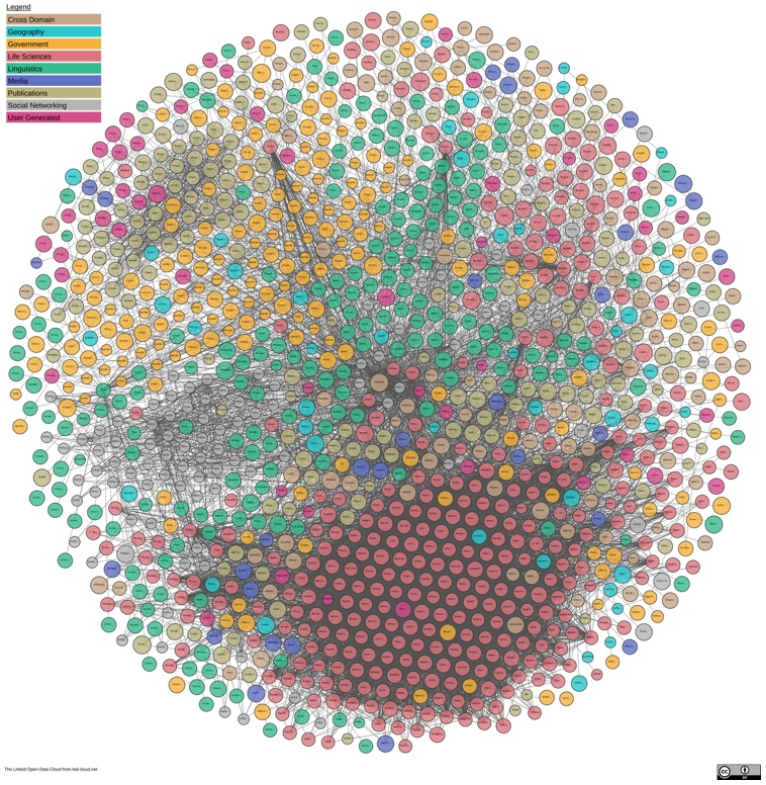
Presentation Outline

- Motivation and Related Works
- Hypothesis for computing true Maximum Cardinalities
- Significant Maximum Cardinality
- Mining Algorithm
- Experiments
- Conclusion

Using Web Knowledge Bases

- **Automatically Generated with**
  - Crowdsourcing
  - Extraction / Integration
- **Used for**
  - Enriching Datasets
  - Linking Datasets in LOD
  - Knowledge Discovery
  - …
- **Need to know when knowledge is completed**

Linked Open Data (2008…



Related Works

- **Interest of information about roles**
  - Data mining on web knowledge bases
  - Characterize query answers

[Darari ISWC'13]
[Galarraga WWW'13]
[Razniewski SIGMOD'15]
[Galarraga WebDB'17]
[Tanon ISWC'17]

- **Mining role cardinality for individual**
  - Exogenous approaches
  - Endogenous approaches (PCA)

[Galarraga WWW'13]
[Galarraga WSDM'17]
[Mirza ISWC'18]

- **Mining role cardinality for concept**
  - Contextual keys
  - Mandatory roles
  - Role minimum/maximum cardinalities

[Pernelle JWS'13]
[Symeonidou ISWC'17]
[Munoz DEXA'17]
[Lajus WWW'18]

A Look on Role Cardinalities in DBpedia

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

Take into account Incorrect Facts

Incorrectness

- About one hundred persons have 2, 3, 4 or 5 declared birth years
- Maximum cardinality
  ≠
  Bigest observed cardinality

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

Another Look on Role Cardinalities in DBpedia

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

Take into account Incompleteness

Incompleteness

- More than half of the persons have only one declared parent
- Maximum cardinality
  ≠
  Most frequently observed cardinality

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

More on Role Cardinalities in DBpedia

$\top$ / 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

Closer Look on Role Cardinalities in DBpedia

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

$\top$ / 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

FootballMatch / team

Take into account Context and Distribution

Context

$\top$ / 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

- Maximum Cardinality is context dependant
- Not always meaningful to compute a Maximum Cardinality

Computing True Maximum Cardinalities

- **If all individuals in KB have only 1 parents?**

- **For M the** true **maximum cardinality of role r**

  - ε, level of incorrectness of a role r: represents the probability to observe an individual having more than M times role r

  - λ, degree of completeness of role r: represents the probability to observe an individual having M times role r

Assumptions: Level of Incorrectness

- **Assumptions for true maximum cardinality computing**
  - The level of incorrectness ε is not significant

ε = 0.43%
0.43% of persons have more than 2 parents

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

Assumptions: Degree of Completeness

- **Assumptions for true maximum cardinality computing**
  - The level of incorrectness ε is not significant
  - The degree of completeness λ is significantly higher

ε = 0.43%

λ = 46.7%
46.7% of persons have 2 parents

For ε = 0.43%, λ greater than 7% is sufficient

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

Likelihood of a Maximum Cardinality i

- Conditional probability $P(X = i \mid X \geq i)$
- **Likelihood** $\tau_i^{C,R}(\mathcal{K}) = \frac{n_i^{C,R}}{n_{\geq i}^{C,R}}$
- Without incorrectness: $P(X = i \mid X \geq i) = 1 \Leftrightarrow$ i is the true maximum

Examples of Likelihood Values

- Likelihood $\tau_i^{C,R}(\mathcal{K}) = \frac{n_i^{C,R}}{n_{\geq i}^{C,R}}$

- $n_{\mathbf{2}}^{\text{Person,parent}}$ = 9,392

- $n_{\geq \mathbf{2}}^{\text{Person,parent}}$ = 9,392+75+9+1
  = 9,477

- $\tau_{\mathbf{2}}^{\text{Person,parent}}(DBpedia) = \frac{\mathbf{9392}}{\mathbf{9477}} = \mathbf{0.991}$

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

Problem

- $\tau_6^{\text{Person,parent}}(DBpedia) = 1$

- Without incorrectness: P(X = i | X ≥ i)=1 ⇔ i is the true maximum card.

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

Correction Using Hoeffding's Inequality

- **Hoeffding's inequality**
  - True for any distribution
  - Upper bound of the deviation when estimating P($X = i \mid X \geq i$)

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

- Given a confidence level $1 - \delta$ we have P($X = i \mid X \geq i$) $\geq \widetilde{\tau}_i(\mathcal{K})$

Examples of Pessimistic Likelihood Values

For $1 - \delta$ = 99%

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

Significant Maximum Cardinality w.r.t. $\mathcal{K}$

- **minimum likelihood threshold** min_τ

- M is a **Significant Maximum Cardinality** iff
  $\mathrm{M} = \arg\max_{i \geq 1} \widetilde{\tau_i}$ and $\widetilde{\tau_M} \geq min_\tau$

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

For $1 - \delta = 99\%$ and $min_\tau = 0.97$ the maximum cardinality of role parent for persons in DBpedia is 2

Algorithm: concept as context

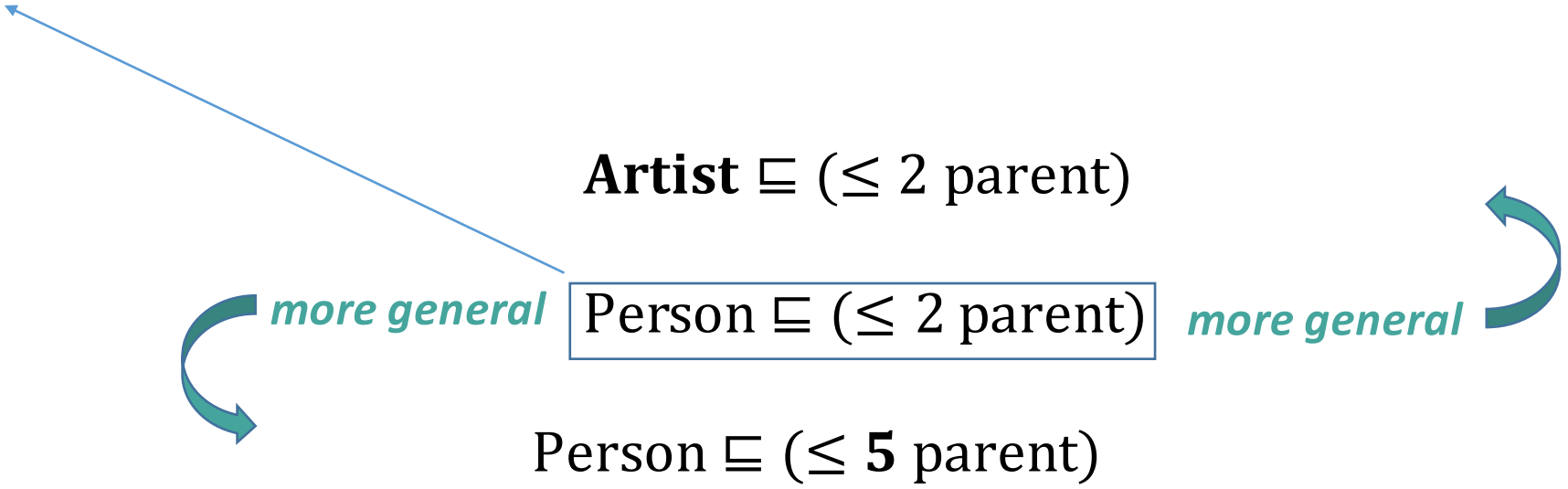
- **Contextual Constraint**
  - Individuals of concept C have at most M role R in $\mathcal{K}$:

$$C \sqsubseteq (\leq M\ R)$$

- **Concept hierarchy:** $\text{Artist} \sqsubseteq \text{Person}$

Minimality

- **Minimal Contextual Constraint**



C3M Algorithm

INPUT

Knowledge base $\mathcal{K}$
Set of roles $\boldsymbol{R}$ in $\mathcal{K}$
Concept hierarchy $(\mathcal{C}, \sqsubseteq)$ in $\mathcal{K}$
Confidence level $1-\delta$
Minimum likelihood threshold min_τ

OUTPUT

Set of all contextual maximum cardinalities

$$\gamma: C \sqsubseteq (\leq M\ R)$$

With $C \in \mathcal{C}$
$R \in \boldsymbol{R}$

Such that γ is minimal w.r.t. $\mathcal{C}$
γ is significant w.r.t. $\mathcal{K}$

On the Web Knowledge Base Scale…

- **DBpedia**
  - ≥ 483,000 concepts
  - ≥ 60,000 roles
- **YAGO**
  - ≥ 500,000 concepts
  - ≥ 93,000 roles
- **…**

First Pruning Criteria: Significant $\widetilde{\tau_M} \geq min_\tau$

- Significant only if there is **enough individuals of C having role R**

  If $|C \sqcap \exists R.\top| < \frac{\log(1/\delta)}{2(1-min_\tau)^2}$ **then no constraint**

  γ: $C' \sqsubseteq (\leq M\ R)$ with $C' \sqsubseteq C$

  can be significant w.r.t. $\mathcal{K}$

  For $1-\delta = 99\%$ and $min_\tau = 0.97$: at least 2558 individuals

Second Pruning Criteria: Minimal

If

$$\gamma_1 : C \sqsubseteq (\leq 1\ R)$$

Then no constraint

$$\gamma_2 : C' \sqsubseteq (\leq M\ R), C' \sqsubseteq C, M \geq 1$$

can be minimal w.r.t. $\mathcal{K}'$s hierarchy of concepts

Experiments

- **All programs and some of the result sets:**
  https://github.com/asoulet/c3m

- **DBpedia**

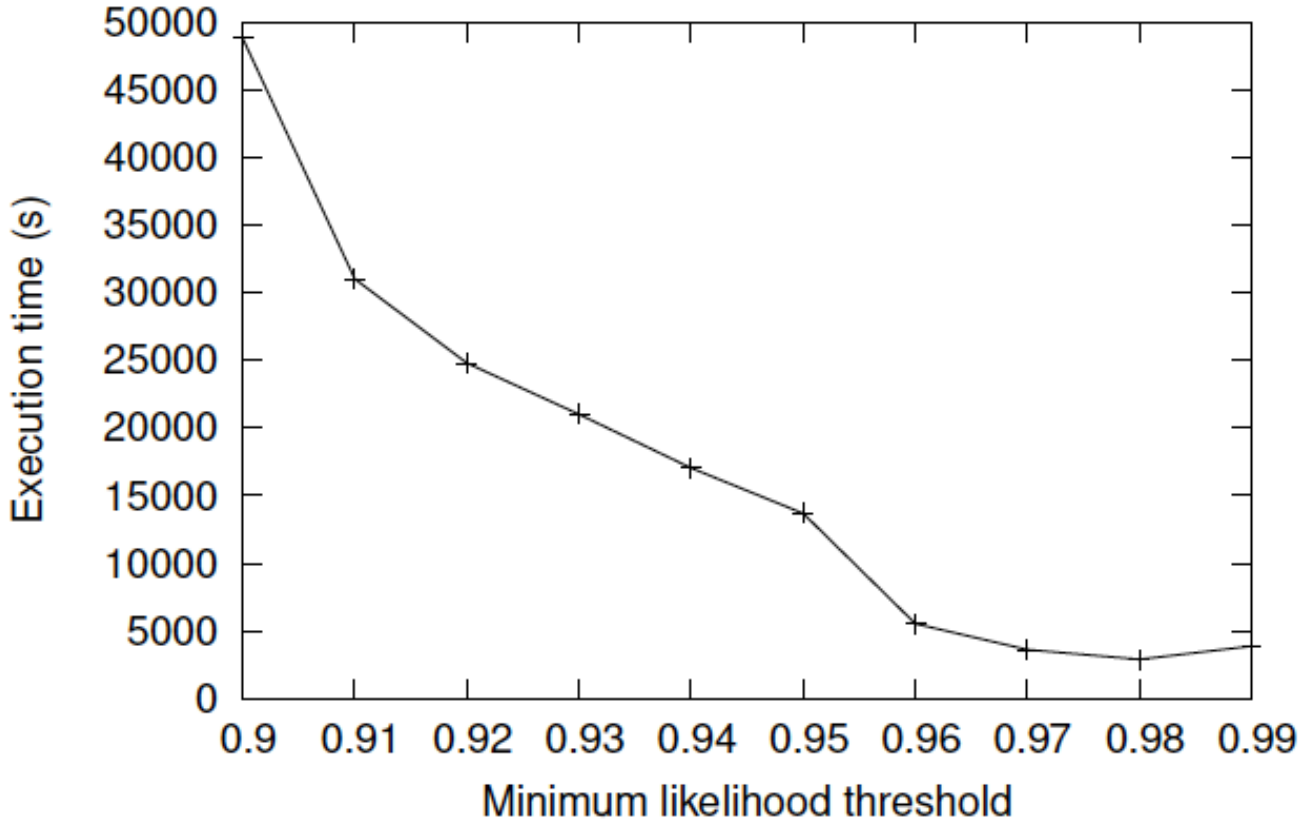
  500 million triples / 480,000 concepts / 60,000 roles

  All experiments done with $1 - \delta = 99\%$

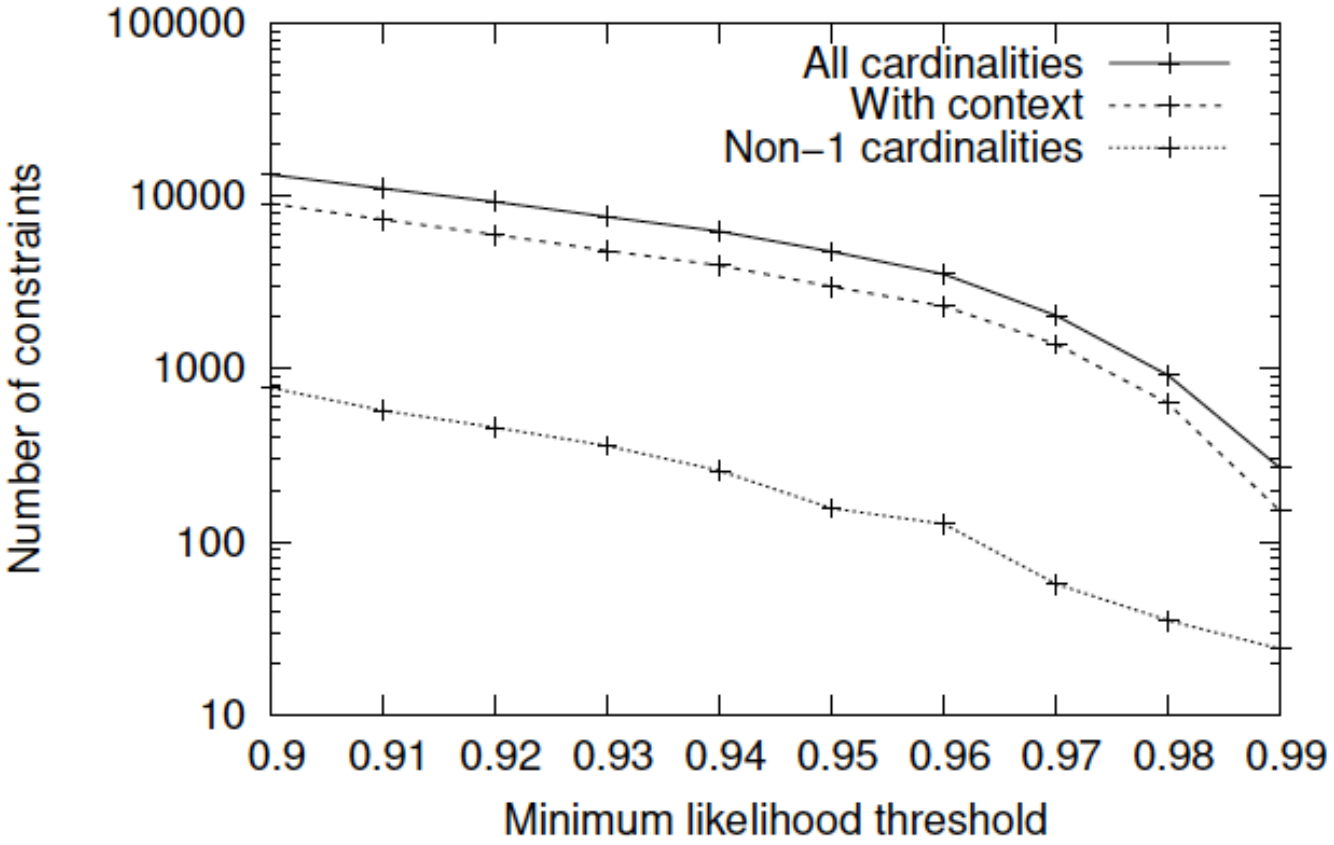
  Queried online via its SPARQL Endpoint

- **TESTED KNOWLEDGE BASES NOT NEED TO BE DOWNLOADED**

Good Scalability



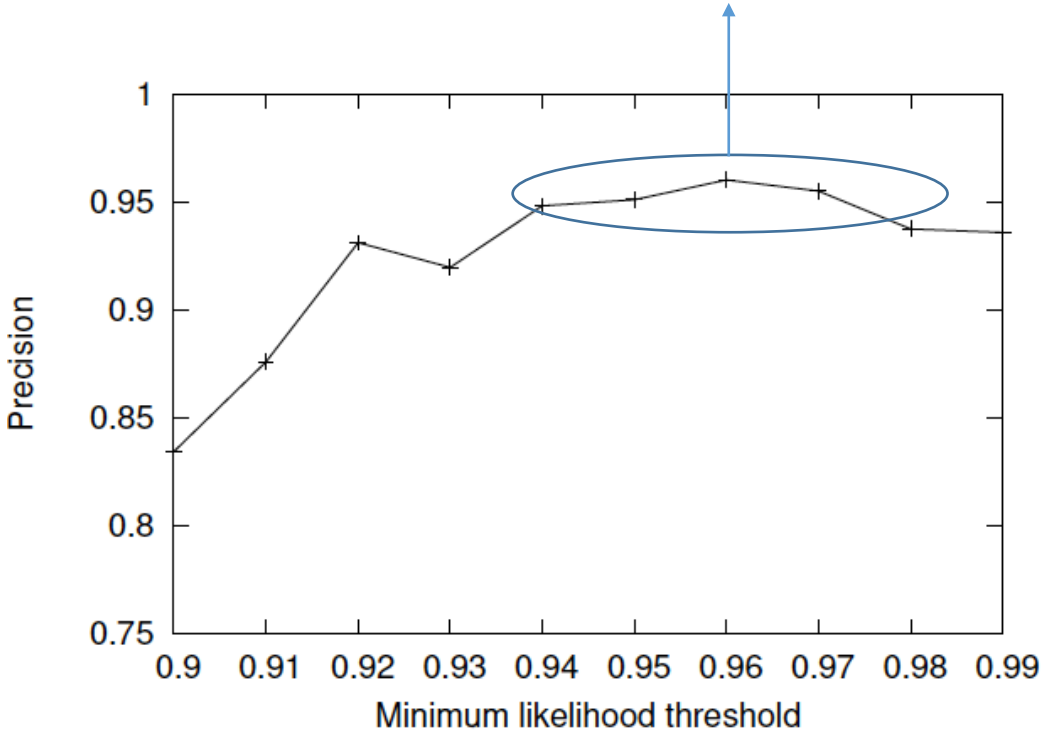
Manageable and interesting constraints



High Precision

- Based on a built ground truth

95% of the
mined constraints are corrects



Conclusion

- **Contributions**
  - A method for computing a Significant maximum cardinality in a Web knowledge base, based on a likelihood measure and Hoeffding's inequality
  - An algorithm for enumerating the set of all Minimal significant maximum cardinalities in a Web knowledge base
- **Program Features**
  - High scalability, high precision, also interesting information about KB content generation
  - No need to download Web knowledge bases
- **Future Work**
  - Compute other Web knowledge base features
  - Use more reasonning (equivalent classes or properties, owl:sameAs, etc.)

Auckland 2019 ISWC

Merci

arnaud.giacometti@univ-tours.fr
beatrice.markhoff@univ-tours.fr
arnaud.soulet@univ-tours.fr

université de TOURS

LIFAT EA 6300
LABORATOIRE D'INFORMATIQUE FONDAMENTALE ET APPLIQUÉE DE TOURS