Using Redescriptions and Association Rules for Mining Definitions in Linked Data

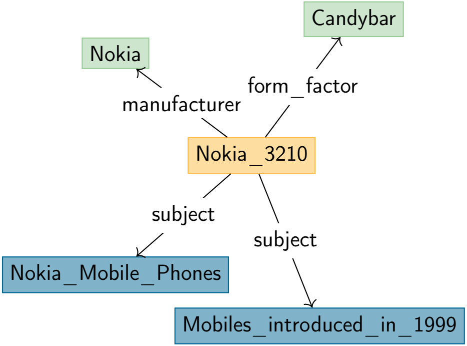
Justine Reynaud, Yannick Toussaint and Amedeo Napoli

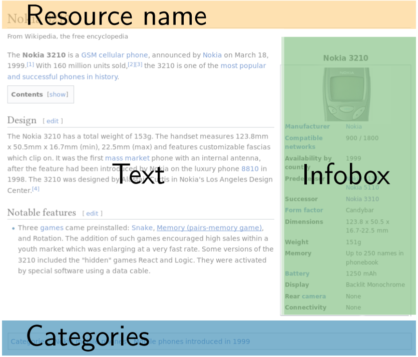
LORIA (Université de Lorraine, INRIA, CNRS), Vandœuvre-les-Nancy, France

Journée thématique EGC & IA : Découverte de connaissances dans le Web de données
10 mai 2019

Introduction — DBpedia

> « *The DBpedia Ontology is a shallow, cross-domain ontology, which has been* ***manually created*** *based on the most commonly used* ***infoboxes within Wikipedia****.* »





Nokia 3210

From Wikipedia, the free encyclopedia

The **Nokia 3210** is a GSM cellular phone, announced by Nokia on March 18, 1999.[1] With 160 million units sold,[2][3] the 3210 is one of the most popular and successful phones in history.

Contents [show]

Design [edit]

The Nokia 3210 has a total weight of 153g. The handset measures 123.8mm x 50.5mm x 16.7mm (min), 22.5mm (max) and features customizable fascias which clip on. It was the first mass market phone with an internal antenna, after the feature had been introduced by Nokia on the luxury phone 8810 in 1998. The 3210 was designed by Alastair Curtis in Nokia's Los Angeles Design Center.[4]

Notable features [edit]

- Three games came preinstalled: Snake, Memory (pairs-memory game), and Rotation. The addition of such games encouraged high sales within a youth market which was enlarging at a very fast rate. Some versions of the 3210 included the "hidden" games React and Logic. They were activated by special software using a data cable.

Nokia 3210

Manufacturer	Nokia
Compatible networks	900 / 1800
Availability by country	1999
Predecessor	[illegible] Nokia 5110
Successor	Nokia 3310
Form factor	Candybar
Dimensions	123.8 x 50.5 x 16.7-22.5 mm
Weight	151g
Memory	Up to 250 names in phonebook
Battery	1250 mAh
Display	Backlit Monochrome
Rear camera	None
Connectivity	None

Categories: Nokia mobile phones | Mobile phones introduced in 1999

Introduction — Problem statement

```
Nancy in France    Paris in France
Nancy in Europe    Paris in Europe
Nancy a City       Paris a City

Rome in Italy      Le_Louvre in France
Rome in Europe     Le_Louvre in Europe
Rome a City        Le_Louvre a Museum
```

$$\text{French\_Cities} = \{\text{Paris}, \text{Nancy}\}$$

How to infer *definitions* in order to *complete* the web of data ?

$$\texttt{French\_Cities} \equiv (\text{a}, \texttt{City}) \sqcap (\texttt{in}, \texttt{France})$$

Data representation

Data representation in FCA

```
Nancy in France   Paris in France
Nancy a City      Paris a City
Rome in Italy     Le_Louvre in France
Rome a City       Le_Louvre a Museum
```

	(in, France) 🇫🇷	(in, Italy) 🇮🇹	(in, Europe) 🇪🇺	(a, City) 🏙	(a, Museum) 🏺
Nancy	×		×	×	
Rome		×	×	×	
Paris	×		×	×	
Le_Louvre	×		×		×

Data representation in FCA

```
Nancy in France
Nancy a City
Rome in Italy
Rome a City
```

```
Paris in France
Paris a City
Le_Louvre in France
Le_Louvre a Museum
```

French_Cities = {Paris, Nancy}
Museums_in_Paris = {Le_Louvre}
European_Capital = {Paris, Rome}

	(in, France) 🇫🇷	(in, Italy) 🇮🇹	(in, Europe) 🇪🇺	(a, City) 🏙️	(a, Museum) 🏺	French_Cities FC	Museums_in_Paris MP	European_Capital EC
Nancy	×		×	×		×		
Rome		×	×	×				×
Paris	×		×	×		×		×
Le_Louvre	×		×		×		×	

Derivation operators and concepts

	🇫🇷	🇮🇹	🇪🇺	🏙️	🏺	FC	MP	EC
Nancy	×		×	×		×		
Rome		×	×	×				×
Paris	×		×	×		×		×
Le_Louvre	×		×		×		×	

{Nancy}' = {🇫🇷, 🇪🇺, 🏙️, FC}

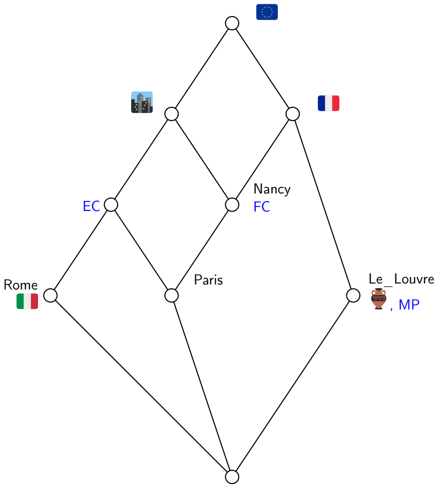
{🇫🇷, 🏙️}' = {Nancy, Paris}

Concept

Given $A \subseteq G$ and $B \subseteq M$, the pair (A, B) is a concept if $A' = B$ and $B' = A$.

({Nancy, Paris}, { 🇫🇷, 🇪🇺, 🏙️, FC }) is a concept.

Concept lattice



Concepts are partially ordered.

- Implication:

$$(in, France) \Rightarrow (in, Europe)$$

- Definition:

$$(a, Museum) \Leftrightarrow MP$$

- Association rule:

$$(in, France) \rightarrow (a, City)$$

Association rules and redescriptions

Association rules

- Searching for dependencies between sets of attributes
- Quality metrics based on confidence

$$conf(X \rightarrow Y) = \frac{\mid X' \cap Y' \mid}{\mid X' \mid}$$

Association rules

- Searching for dependencies between sets of attributes
- Quality metrics based on confidence

$$conf(X \to Y) = \frac{\mid X' \cap Y' \mid}{\mid X' \mid}$$

Example

	🇫🇷	🇮🇹	🇪🇺	🏙️	🏺	FC	MP	EC
Nancy	×		×	×		×		
Rome		×	×	×				×
Paris	×		×	×		×		×
Le_Louvre	×		×		×		×	

$$conf(\{\text{🇫🇷}\} \to \{\text{🏙️}\}) = \frac{\mid \text{🇫🇷}' \cap \text{🏙️}' \mid}{\mid \text{🇫🇷}' \mid} = \frac{\mid \{Nancy, Paris\} \mid}{\mid Nancy, Paris, Le\_Louvre \mid} = \frac{2}{3}$$

Association rules – Eclat [Zaki, 2000]

- Exhaustive enumeration
- Rules are unidirectional
- Post-processing in order to select rules satisfying criteria

Quasi-definition

A quasi-definition $X \leftrightarrow Y$ holds with a confidence θ iff

$$min(conf(X \rightarrow Y), conf(Y \rightarrow X)) = \theta$$

Redescriptions – ReReMi [Galbrun and Miettinen, 2012]

- Searching for two sets of attributes that occurs in the same objects
- Rules are bidirectional and more expressive than association rules
- Quality metrics based on Jaccard coefficient

$$Jacc(X \leftrightarrow Y) = \frac{\mid X' \cap Y' \mid}{\mid X' \cup Y' \mid}$$

Redescriptions – ReReMi [Galbrun and Miettinen, 2012]

- Searching for two sets of attributes that occurs in the same objects
- Rules are bidirectional and more expressive than association rules
- Quality metrics based on Jaccard coefficient

$$Jacc(X \leftrightarrow Y) = \frac{\mid X' \cap Y' \mid}{\mid X' \cup Y' \mid}$$

Example

	🇫🇷	🇮🇹	🇪🇺	🏙️	🏺	FC	MP	EC
Nancy	×		×	×		×		
Rome		×	×	×				×
Paris	×		×	×		×		×
Le_Louvre	×		×		×		×	

{🇫🇷} ↔ {FC} $Jacc(\{\text{🇫🇷}\} \leftrightarrow \{FC\}) = \frac{2}{3}$

Redescriptions – ReReMi [Galbrun and Miettinen, 2012]

- Searching for two sets of attributes that occurs in the same objects
- Rules are bidirectional and more expressive than association rules
- Quality metrics based on Jaccard coefficient

$$Jacc(X \leftrightarrow Y) = \frac{\mid X' \cap Y' \mid}{\mid X' \cup Y' \mid}$$

Example

	🇫🇷	🇮🇹	🇪🇺	🏙️	🏺	FC	MP	EC
Nancy	×		×	×		×		
Rome		×	×	×				×
Paris	×		×	×		×		×
Le_Louvre	×		×		×		×	

{🇫🇷, 🏙️} ↔ {FC}   *Jacc*({🇫🇷, 🏙️} ↔ {*FC*}) $= \frac{2}{2} = 1$

Experiments

Experiments : Datasets

- Datasets extracted from DBpedia, thanks to a SPARQL query
- Various sizes and domains

	Person	Object	Film
Small	Turing_Award	Samsung_Galaxy	Hospital_films
Medium	Women_Mathematicians	Smartphones Sports_cars	Road_movies
Large	Mathematicians	—	French_films

Experiments : Datasets (statistics)

Dataset	Triples	$\|G\|$	$\|M\|$	$\|M_{subj}\|$	$\|M_{descr}\|$	$\|P\|$	δ
Samsung_Galaxy	940	59	277	30	247	33	5.2e−2
Turing_Award_laureates	2642	65	1360	503	857	35	2.2e−2
Hospital_films	1984	71	1265	490	775	46	1.6e−2
Women_mathematicians	9652	552	4243	1776	2467	98	2.9e−3
Smartphones	8418	598	2089	359	1730	98	5.8e−3
Sports_cars	9047	604	2730	435	2295	61	4.7e−3
Road_movies	20056	689	9314	2652	6662	103	2.4e−3
Mathematicians	32536	1660	12279	3848	8431	202	1.2e−3
French_films	121496	6039	25487	6028	19459	111	6.4e−4

Experiments: extracted rules — `Turing_award_laureates`

Association Rules

$$\begin{aligned}
\textbf{Harvard\_University\_alumni} &\equiv \exists \text{almaMater.Harvard\_University} \sqcap \text{Agent} \sqcap \text{Person} \\
&\quad \sqcap \text{Scientist} \\
\textbf{Harvard\_University\_alumni} &\equiv \exists \text{almaMater.Harvard\_University} \sqcap \\
&\quad \exists \text{award.Turing\_Award} \sqcap \text{Agent} \sqcap \text{Person} \sqcap \text{Scientist} \\
\textbf{National\_Medal\_of\_Science\_l} &\equiv \exists \text{award.National\_Medal\_of\_Science} \sqcap \text{Agent} \sqcap \text{Person} \\
&\quad \sqcap \text{Scientist} \\
\textbf{M.\_I.\_T.\_faculty} &\not\equiv \exists \text{award.Turing\_Award} \sqcap \text{Agent} \sqcap \text{Person} \\
&\quad \sqcap \exists \text{birthPlace.New\_York\_City}
\end{aligned}$$

Redescriptions

$$\begin{aligned}
\textbf{Harvard\_University\_alumni} &\equiv \exists \text{almaMater.Harvard\_University} \\
\textbf{Stanford\_University\_alumni} &\equiv \exists \text{almaMater.Stanford\_University} \\
\textbf{National\_Medal\_of\_Science\_l.} &\equiv \exists \text{award.National\_Medal\_of\_Science} \\
\textbf{British\_computer\_scientists} &\not\equiv \exists \text{award.Fellow\_of\_the\_Royal\_Society}
\end{aligned}$$

Experiments: extracted rules — Smartphones

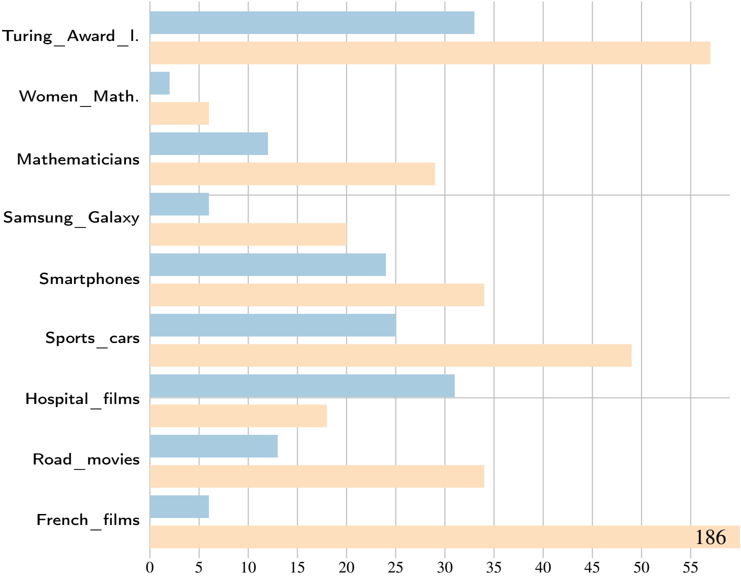
Association Rules

$$\begin{aligned}
\textbf{Nokia\_mobile\_phones} &\equiv \exists\text{manufacturer.Nokia} \sqcap \text{Device}\\
\textbf{Samsung\_Galaxy} &\equiv \exists\text{manufacturer.Samsung\_Electronics} \sqcap \text{Smartphones}\\
&\quad \sqcap \text{Device}\\
\textbf{Mobile\_operating\_systems} &\equiv \text{Software} \sqcap \text{Work}\\
\textbf{Sony\_mobile\_phones} &\not\equiv \exists\text{input.Capacitive\_sensing} \sqcap \exists\text{input.Proximity\_sensor}\\
&\quad \sqcap \exists\text{input.Touchscreen}
\end{aligned}$$

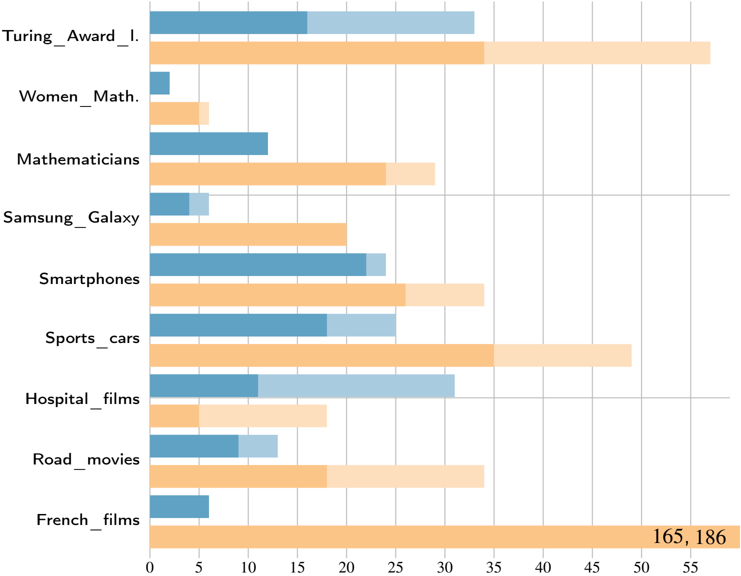
Redescriptions

$$\begin{aligned}
\textbf{Nokia\_mobile\_phones} &\equiv \exists\text{manufacturer.Nokia}\\
\textbf{Samsung\_Galaxy} &\equiv \exists\text{manufacturer.Samsung\_Electronics}\\
&\quad \sqcap \exists\text{operatingSystem.Android\_OS}\\
\textbf{Mobile\_operating\_systems} &\equiv \text{Software} \sqcap \text{Work}\\
\textbf{MeeGo\_Devices} &\not\equiv \exists\text{operatingSystem.Sailfish\_OS}
\end{aligned}$$

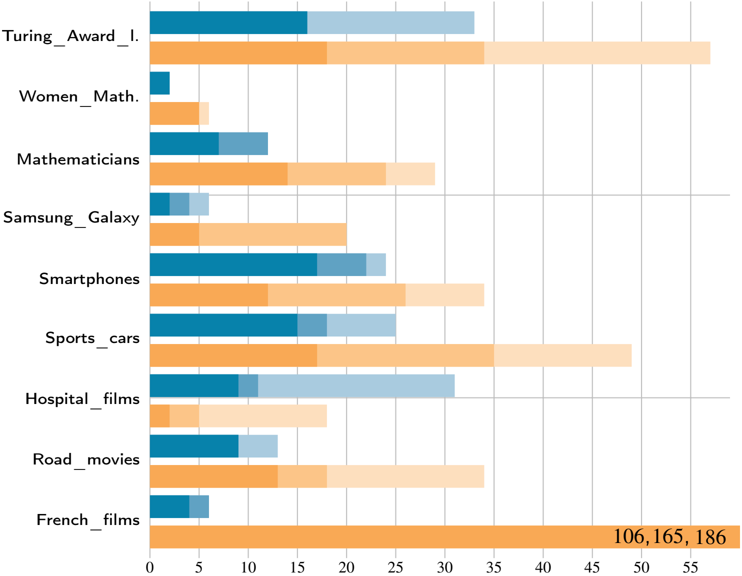
Experiments : Rules extracted (statistics)



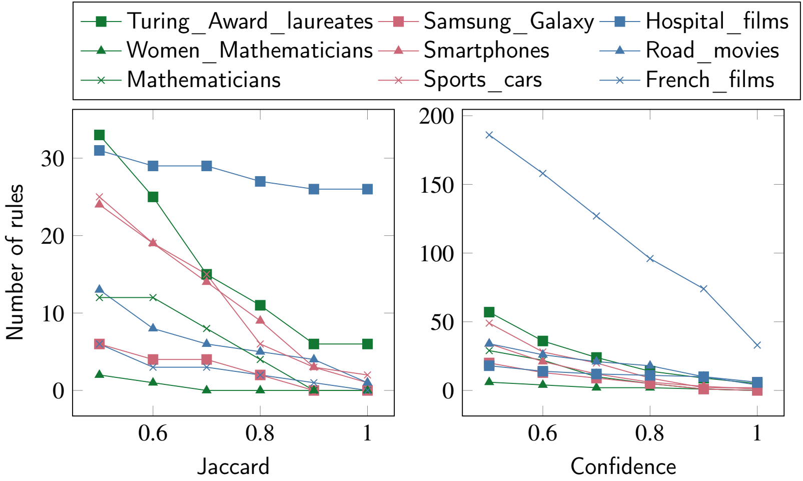
Experiments : Rules extracted (statistics)



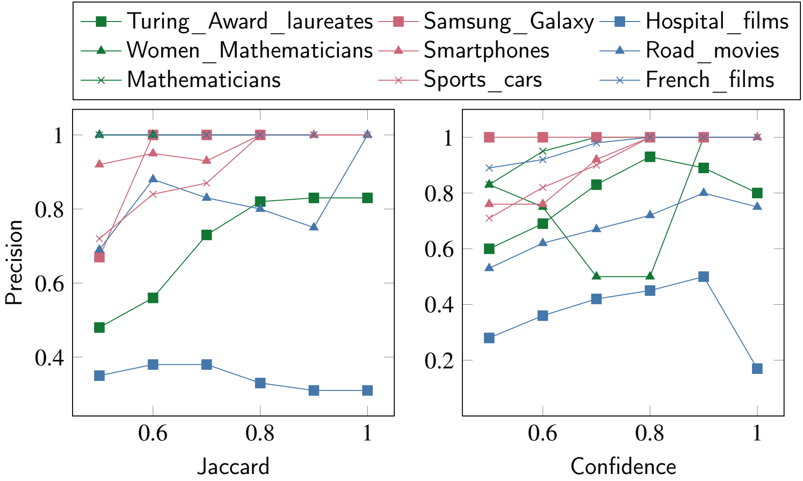
Experiments : Rules extracted (statistics)



Experiments – Results

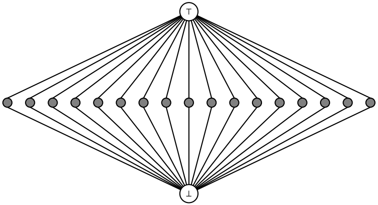


Experiments – Results

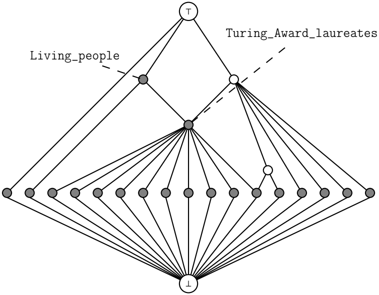


Discussion – Ordering categories

Concept lattices of the defined categories (`Turing_Award_laureates`).



Redescriptions



Association rules

Definition of `Harvard_University_alumni`

Red. ∃almaMater.Harvard_University

A.R. ∃almaMater.Harvard_University ⊓ Agent ⊓ Person ⊓ Scientist

Discussion – Predicates

	Pred.	Pred (D_{RD})	Pred (D_{QD})
Turing_Award_laureates	35	4	5
Women_Mathematicians	98	2	3
Mathematicians	202	4	5
Samsung_Galaxy	33	2	7
Smartphones	98	5	8
Sports_cars	61	3	5
Hospital_films	46	5	2
Road_movies	103	3	7
French_films	111	4	10

dbo:award (261), ***rdf:type (186)***, dbo:knownFor (182), dbo:doctoralStudent (148), **dbp:workInstitution (123)**, dbo:birthPlace (117), **dbo:almaMater (110), dbo:field (84)**, dbo:doctoralAdvisor (36), dbo:deathPlace (36), dbp:workplaces (28), dbp:workInstitutions (26), dbo:influenced (23), dbo:nationality (15)

Conclusion and Future work

- Redescriptions interesting for defining categories
- Association Rules and Redescriptions complete each other

- *Are definitions operational ?*
  Integration to knowledge base.
- *Can we (should we) use more expressive definitions ?*
  $C \equiv A \sqcup B$ or $C \equiv \neg A$

Thanks for your attention.
Questions ?

justine.reynaud@loria.fr

Galbrun, E. and Miettinen, P. (2012).
From Black and White to Full Color: Extending Redescription Mining Outside the Boolean World.
Statistical Analysis and Data Mining, 5(4):284–303.

van Leeuwen, M. and Galbrun, E. (2015).
Association Discovery in Two-View Data.
TKDE, 27(12):3190–3202.

Zaki, M. J. (2000).
Scalable algorithms for association mining.
TKDE, 12(3):372–390.