UNIVERSITÄT Mannheim

Linked Open Data enhanced Knowledge Discovery

Introducing the RapidMiner Linked Open Data Extension

The Web is Full of Data...



09/30/15

Motivating Example

Understanding population changes in the Netherlands



09/30/15

Motivating Example

- Understanding population changes in the Netherlands
- What we can see in the data
 - population changes by municipality are very diverse
 - ranging from -12% to +53% over the last 15 years
- · What we cannot see from the data
 - How do growing regions differ from shrinking ones?
 - Which factors drive people's movements?
- As very often, we need more knowledge...

Aarlanderver	en				A
Enkhuizen	Municipality	17,033	17,804	18,173	18,347
Haarlem	Municipality	148,373	147,640	150,670	156,660
Haarlemmerliede en Spaarnwoude	Municipality	5,578	5,429	5, <mark>4</mark> 32	5,573
Haarlemmermeer	Municipality	113,543	140,648	143,374	144,182
Heemskerk	Municipality	36,068	38,381	39,206	39,139
Heemstede	Municipality	26,003	25,626	26,297	26,467
Heerhugowaard	Municipality	45,551	50,390	51,985	53,548
Heiloo	Municipality	21,897	22,024	22,580	22,548
Hilversum	Municipality	82,769	83,815	84,984	87,230

Motivating Example

- Proposed approach:
 - link data at hand to LOD Cloud
 - harvest additional information about regions
 - look for interesting patterns

Aarlanderveen		_			A
Enkhuizen	Municipality	17,033	17,804	18,173	18,347
Haarlem	Municipality	148,373	147,640	150,670	156,660
Haarlemmerliede en Spaarnwoude	Municipality	5,578	5,429	5,432	5,573
Haarlemmermeer	Municipality	113,543	140,648	143,374	144,182
Heemskerk	Municipality	36,068	38,381	39,206	39,139
Heemstede	Municipality	26,003	25,626	26,297	26,467
Heerhugowaard	Municipality	45,551	50,390	51,985	53,548
Heiloo	Municipality	21,897	22,024	22,580	22,548
Hilversum	Municipality	82,769	83,815	84,984	87,230



RapidMiner Linked Open Data Extension

Introducing RapidMiner:

- An open source platform for data mining and predictive analytics
- Processes are designed by wiring operators in a GUI (no programming)
- Operators for data loading, transformation, modeling, visualization, ...
- Scalable, distributed, parallel processing in a cloud environment
- 200,000 active users

09/30/15



• Developers can write their own *extensions*

RapidMiner Linked Open Data Extension

- The extension adds operators for
 - accessing local and remote (Linked and non Linked) data
 - linking local to remote data
 - combining data from various sources
 - automatically following links to other datasets
- Data analysts can use it without knowing RDF, SPARQL, etc.



09/30/15

Example Use Case

- Understanding population changes in the Netherlands
- RapidMiner workflow:
 - Import original table
 - Link municipalities to DBpedia
 - alternative: link provinces to Eurostat
 - Build enriched table
 - Analyze the results

Aarlanderveen	~				4
Enkhuizen	Municipality	17,033	17,804	18,173	18,347
Haarlem	Municipality	148,373	147,640	150,670	156,660
Haarlemmerliede en Spaarnwoude	Municipality	5,578	5,429	5,432	5,573
Haarlemmermeer	Municipality	113,543	140,648	143,374	144,182
Heemskerk	Municipality	36,068	38,381	39,206	39,139
Heemstede	Municipality	26,003	25,626	26,297	26,467
Heerhugowaard	Municipality	45,551	50,390	51,985	53,548
Heiloo	Municipality	21,897	22,024	22,580	22,548
Hilversum	Municipality	82,769	83,815	84,984	87,230

Example Findings

- Growing regions: Flevoland, Utrecht, North/South Holland
- Shrinking regions: Limburg, Groningen, Friesland
- Provincal capitals are growing
- Growth in regions with high population
- Growth in regions with high income
 - but also: growth in regions with high unemployment

Example Findings

• Negative correlation between growth and elevation?!



09/30/15

Behind the Scenes: RapidMiner LOD Extension

• RapidMiner uses a tabular data model

			h	nput	Tabl	е				Link		5	Addi	tiona	al Att	ribut	es		
Row No.	cylinders	displacement	t horsepower	weight	accelaeration	model	origin	mpg	car	car_uri	http://dbped	http://dbped.	http://dbped	http://dbped.	http://dbped	http://dbped	http://dbped	http://dbped	. ht
1	8	307	130	3504	12	70	1	18	chevrolet che	http://dbpedi	0	0	0	1	1	0	0	0	0 ^
2	8	350	165	3693	11.500	70	1	15	buick skylark	http://dbpedi	1	0	0	1	1	0	0	0	0
3	8	318	150	3436	11	70	1	18	plymouth sat	http://dbpedi	1	0	0	1	1	0	0	0	0
4	8	304	150	3433	12	70	1	16	amc rebel	http://dbpedi	1	0	0	1	1	0	0	0	0
5	8	302	140	3449	10.500	70	1	17	ford torino	http://dbpedi	1	0	0	1	0	0	0	0	0
6	8	429	198	4341	10	70	1	15	ford galaxie	http://dbpedi	1	0	0	1	1	0	0	0	0
7	8	454	220	4354	9	70	1	14	chevrolet imp	http://dbpedi	1	0	0	1	1	0	0	0	0
8	8	440	215	4312	8.500	70	1	14	plymouth fur	http://dbpedi	1	0	0	1	1	0	0	0	0
9	8	455	225	4425	10	70	1	14	pontiac catal	http://dbpedi	1	1	0	1	1	0	0	0	0
10	8	390	190	3850	8.500	70	1	15	amc ambas:	http://dbpedi	1	0	0	1	1	0	0	0	0
11	8	383	170	3563	10	70	1	15	dodge challe	http://dbpedi	1	0	0	1	1	0	0	0	0
12	8	340	160	3609	8	70	1	14	plymouth 'cu	http://dbpedi	1	0	0	1	1	0	0	0	0
13	8	400	150	3761	9.500	70	1	15	chevrolet mc	http://dbpedi	1	0	0	1	1	0	0	0	0
14	8	455	225	3086	10	70	1	14	buick estate	http://dbpedi	1	0	0	1	0	0	0	0	0
15	4	113	95	2372	15	70	3	24	toyota coron;	http://dbpedi	1	0	0	1	1	0	0	0	0
16	6	198	95	2833	15.500	70	1	22	plymouth du	http://dbpedi	1	0	0	1	0	0	0	0	0
17	6	199	97	2774	15.500	70	1	18	amc hornet	http://dbpedi	1	0	0	1	1	0	0	0	0
18	6	200	85	2587	16	70	1	21	ford maveric	http://dbpedi	1	0	0	1	0	0	0	0	0
19	4	97	88	2130	14.500	70	3	27	datsun	http://dbpedi	0	0	0	1	0	0	0	0	0
20	4	97	46	1835	20.500	70	2	26	volkswagen	http://dbpedi	1	0	0	1	0	1	0	0	0
21	4	110	87	2672	17.500	70	2	25	peugeot 504	http://dbpedi	0	0	0	0	0	0	0	0	0
22	4	107	90	2430	14.500	70	2	24	audi 100	http://dbpedi	0	0	0	0	0	0	0	0	0
23	4	104	95	2375	17.500	70	2	25	saab 99	http://dbpedi	0	0	0	0	0	0	0	0	0
24	4	121	113	2234	12.500	70	2	26	bmw 2002	http://dbpedi	0	0	0	0	0	0	0	0	0
25	6	199	90	2648	15	70	1	21	amc gremlin	http://dbpedi	1	0	0	1	0	0	0	0	0
26	8	360	215	4615	14	70	1	10	ford f250	http://dbpedi	1	0	0	1	0	0	0	0	0
27	8	307	200	4376	15	70	1	10	chevy	http://dbpedi	1	0	0	1	1	0	0	0	0
28	8	318	210	4382	13.500	70	1	11	dodge d	http://dbpedi	0	0	0	1	0	0	0	0	0
<																		-	2

09/30/15

Behind the Scenes: RapidMiner LOD Extension

- Linking local data to LOD Sources
 - based on URI patterns
 - based on text search
 - using specialized services (e.g., DBpedia Lookup)
- Following links
 - e.g., automatically follow all owl:sameAs links to other datasets to a certain depth
- Harvesting attributes
 - e.g., add all numeric attributes found
 - built-in support for aggregations



Web Validator

Exa

Att

Pattern-based...

Att

Exa

Behind the Scenes: RapidMiner LOD Extension

- Matching and fusion
 - e.g., many sources contain "population" as an attribute
 - automatic identification of similar attributes
 - automatic fusion using different policies
- Attribute set filtering
 - exploiting schema information
 - more effective in finding redundant attributes



Full RapidMiner Workflow for the Example



09/30/15

- Analyzing unemployment in France (SemStats'13)
 - using background knowledge from DBpedia, Eurostat, Linked Geo Data
 - exploiting links from DBpedia to GADM for visualization



(a) Unemployment by region



(b) Heat map of police stations

09/30/15

Heiko Paulheim

15

- Example correlations for unemployment in France:
 - African islands, Islands in the Indian Ocean, Outermost regions of the EU (positive)
 - GDP (negative)
 - Disposable income (negative)
 - Hospital beds/inhabitants (negative)
 - RnD spendings (negative)
 - Energy consumption (negative)
 - Population growth (positive)
 - Casualties in traffic accidents (negative)
 - Fast food restaurants (positive)
 - Police stations (positive)

- Data Set: Suicide rates by country
 - http://www.washingtonpost.com/wp-srv/world/suiciderate.html
- Findings for suicide rates
 - Democracies have lower suicide rates than other forms of government
 - High HDI \rightarrow low suicide rate
 - High population density \rightarrow high suicide rate
 - By geography:

09/30/15

- At the sea \rightarrow low
- In the mountains \rightarrow high
- High Gini index \rightarrow low suicide rate
 - High Gini index ↔ unequal distribution of wealth
- High usage of nuclear power \rightarrow high suicide rates

- Data set: Durex worldwise survey on sexual activity
 - http://chartsbin.com/view/uya
- Findings:
 - By geography:
 - High in Europe, low in Asia
 - Low in Island states
 - By language:
 - English speaking: low
 - French speaking: high
 - Low average age \rightarrow high activity
 - High GDP per capita \rightarrow low activity
 - High unemployment rate \rightarrow high activity
 - High number of ISP providers \rightarrow low activity

Caveat

• We have only been analyzing *correlations* here.



• Incident detection from Twitter



09/30/15



- Using LOD+RapidMiner
 - automatically learns a model
 - avoids overfitting

- Building Semantic Recommeder Systems (ESWC'14)
- Combines two extensions:
 - Linked Open Data extension
 - Recommender system extension
- Use data about books for content-based recommender
 - best system (out of 24) on two out of three tasks
 - used data from DBpedia and RDF Book Mashup



09/30/15

- Compare Hilversum to other Cities in the Netherlands
 - find distinctive features
- Finding the needles in the haystack of statements about Hilversum in DBpedia

Inder Schen Lander 1999 (1999) Anatol (1999) Anat	
Parting A. A	
Analysis • 10 Analysis	
	And a second
An alter of the second secon	state Section S
Abore of the second sec	
الم	
لن استعلی است به است	
ن هندی است به است به است به	
 An example of a second s	
ن به	
هند استهدا هی به می به می به هی ب به هی به ی به هی	areau careau care of specification area of s
Antipic Antipic Section Antipic Section Antipic Section	
	An A Serie A Serie A Series A
Ans. Information (Control of Control of Co	
	հատկաս՝ նատկ նատկ նատկ նատկ նատկ նատկ նատկ նատկ
 An of Segmentation (Segmentation (Segmentatio	ii
	սութան Դու Դու Դու Դու Դու Դու Դու Դու Դու Դու
	Your Your Your Your Your Your Your Your
	Here Services Service
	Mail Sec. Sec. <td< td=""></td<>
Appendix and appendix app	ee_jaks/aks =
	۵۲ ۵۳۵٬۵۳۹ ۵۳ ۵۳۵٬۵۳۹ ۱۳۵ ۵۳۵ ۱۳۵ ۵۳ ۱۳۵ ۵۲ ۱۳۵ ۵۲ ۱۹۹ ۱۳۵ ۲۰ ۱۹
	ամ հայն ամ հայն ամ հայուն ամ
School (Section	առների նառներ հարչը հատ հարչը հատ հատ հարչը հատ հատ հատ հատ հատ հատ հատ հատ
An per-period	
A) Sampling A) Sampling A) Sampling B) Sampling A) Sampling	
	الله الله الله الله الله الله الله الله
	Դու Սանգարան հետ չո հարտերական հետ չուն առաջություն առաջություն Դուր չուն Դուր չուն
Action of a second	នេះកាន់កែវ (ឆ្ល នៅស្តេ) នៅស្តេ) កំពុង (ឆ្លាំនាក់ សក្សមកម្មនេះ (សកមមម)
	an Tapi Mangina Yang Jana
이 있는 것을 가 있는 것을 하는 것을 수가 있다. 것을 하는 것을 하는 것을 하는 것을 하는 것을 수가 있다. 것을 수가 있는 것을 하는 것을 수가 있는 것을 수가 있는 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 수가 있는 것을 수가 있는 것을 수가 있는 것을 수가 있다. 것을 것 같이 않는 것 같이 않는 것 같이 않다. 것을 수가 있는 것 같이 않다. 것을 수가 있는 것 같이 않는 것 같이 않다. 것을 수가 않아야 않아요. 것을 수가 있다. 것을 수가 같이 않아야 않아. 것을 수가 있는 것 같이 않아. 것을 수가 있는 것 같이 않아. 것을 수가 같이 않아. 것을 수가 있다. 것을 수가 것 같이 않아. 것 같이 않이	an mayo Magania ang Tang
န <table-cell> လာနားလာနားလာနားလာနားလာနားလာနားလာနားလ</table-cell>	fen (len (n ang / ng) (kenten) Lennant
ութատրություն ու	ana (a) (ana ini a)
	and the second se
- Ag hangan	in an
ծնչեպում՝ է Շյո մուրազեն է Շրջանն Անդապես Հետում Անդապես Հետում Անդատես է	
نوْسِتَوْرَكُ نَوْسِتَوَرْنُ مَنْ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ نَوْسِتَهُ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ مُوْسِتَهُ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ مُوْسِتَهُ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ مُوْسِتَهُ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ مُوْسِتَهُ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ مُوْسِتَهُ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ اللَّذِي اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ عَلَيْهِ عَلَيْهِ مُوْسِتَعْلَيْ اللَّهِ عَلَيْهِ اللَّعَلَيْلِي اللَّهِ عَلَيْهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْهِ اللَّهِ عَلَيْنَةِ اللَّهِ عَلَيْنَةِ اللَّهِ عَلَيْ الْعَلَيْنَا اللَّهِ اللَّذَى اللَّهِ عَلَيْ اللَّهِ عَلَيْلَةِ عَلَيْ عَلَيْلَةِ عَلَيْ اللَّهِ عَلَيْلَةِ عَلَيْ اللَّهِ عَلَيْ اللَّالِي اللَّهِ عَلَيْلَةِ عَلَيْ	ad Happ)
ն ինչ որվերան՝ ն հետում ինչ որսեն հետում է հետում ինչ որսենան՝ հետում է հետում ինչ որսենան՝ հետում է հետում ինչ որսենան՝ հետում է հետում որսենան՝ հետում է հետում հետում է հետում հետում է հետում հետում է հետում հետում է հետում հետում է հետում	lighted getter of Historium in a multiple reg of Nath Helland
a da pertana a la da	
ing minani ani di Angi Ang minani ani di Angi Ang papalain Caranya Maka Ang pani Caranya Angi Ang mini Caranya Angi Ang mini Caranya Angi Ang mahalan Angi Ang mahalan Angi	
ag na se anno 1995 a	Autor
ng paparan ng panalan ng pan	
ing pantal Cash Type a dark dag satularan Type a dark dag satularan Type a dark dag satularan Type a dark dag satularisan a dark dag satularisan a dark	autore
deg antikerner (type et de L deg antikelt in et de L deg antikerner et de L de L	The second se
Agarháilte (2) Agasth (2) Agasth (2) Agastháite Agasthá	a y juni y la juni a su a s
ng salah kalarkana dari Ang salah kalarkana dari Ang salah kalarkana dari	jadirlage)
in a labor to	Contract.
	and generating global
101 march 101	terrengy of sector
ing image and a start	ins Juge June Ju
ing with a Mate	• • • • • • • • • • • • • • • • • • • •
ing de Dhai e +1	
ng wohalou •d	
	upanan 1911 Juan Danis ana
- 400 - 400 - 400 - 400 - 400 - 400 - 400	Sale_period: Sale_period: Sale_period: Sale_period: Sale_period:

- Compare Hilversum to other Cities in the Netherlands
 - find distinctive features

Top Facts for Hilversum compared to the entities of the class Cities In The Netherlands:

ntries	Search:
tatements	Rules
	Entities of type CitiesInTheNetherlands usually have isPartOf, but Hilversum doesn't have!
	Entities of type CitiesInTheNetherlands usually don't have hometown, but Hilversum has!
	Entities of type CitiesInTheNetherlands usually are not of type OlympicModernPentathIonVenues, but Hilversum is!
	Entities of type CitiesInTheNetherlands usually don't have recordedIn, but Hilversum has!
	Entities of type CitiesInTheNetherlands usually are not of type OlympicEquestrianVenues, but Hilversum is!
	Entities of type CitiesInTheNetherlands usually don't have headquarter, but Hilversum has!
	Entities of type CitiesInTheNetherlands usually don't have location, but Hilversum has!
	Entities of type CitiesInTheNetherlands usually don't have ground, but Hilversum has!
	Entities of type CitiesInTheNetherlands usually are not of type PopulatedPlacesInNorthHolland, but Hilversum is!
	Entities of type CitiesInTheNetherlands usually are not of type 1928SummerOlympicVenues, but Hilversum is!
t	itries iatements

Showing 1 to 10 of 10 entries

Previous 1 Next

09/30/15

- Compare Hilversum to other Cities in the Netherlands
 - find distinctive features
- TopFacts application
 - Demonstration at ISWC 2015
 - Combines Linked Open Data with attribute-wise outlier detection [see Paulheim/Meusel, Machine Learning 100(2-3), 2015]

- Compare Hilversum to other Cities in the Netherlands
 - find distinctive features
- Hilversum is
 - a city where the modern Pentathlon olympics have been held
 - the headquarter of many media companies
 - a place where many music recordings have been made







09/30/15

- Debugging Linked Open Data
 - loading a set of statements
 - augment with additional features
 - run outlier detection
 - again: a special extension
- Example: identify wrong dataset interlinks (WoDOOM'14)
 - AUC up to 85%



09/30/15

Summary

- The RapidMiner LOD Extension
 - brings data analysis to the web of data
 - can be used by data analysts without learning SPARQL
- Availability
 - on the RapidMiner marketplace
 - installable from inside RapidMiner
 - >9,000 installations and counting



09/30/15

Take Home Messages

- The Web is full of data
 - …and more and more becomes Linked Data
- Intelligent data processing
 - helps unlocking the potential of that data
 - enables intelligent applications
- A good fit
 - Sophisticated analytics platforms (e.g., RapidMiner), and
 - Linked Open Data



09/30/15

Feedback?

Heiko Paulheim

heiko@informatik.uni-mannheim.de @heikopaulheim

09/30/15

UNIVERSITÄT Mannheim

Linked Open Data enhanced Knowledge Discovery

Introducing the RapidMiner Linked Open Data Extension