Building Knowledge-Intensive Applications with Linked Open Data*

*using examples from the domains sex, drugs, and crime

Chris  Bizer, Heiko Paulheim, University of Mannheim
1982

"We are drowning in data, but starving for knowledge"

John Naisbitt
Data vs. Knowledge

• There is plenty of data
  – Linked (Open) Data
  – Government Data
  – Sensor Data
  – Social Networks
  – ...

• ...but data is not knowledge

• Knowledge is required for
  – Assessments
  – Actions
Crime Data Live on the Web
Crime Data Live on the Web

• data.seattle.gov:
  – Live 911 call data
  – Open for mashups

• Problem for the response team:
  – Quick decisions required
    • With severe effects...
  – Minimal information
  – Missing background knowledge
    • ...but lots of potential sources
There's a Fire!

- **Knowledge:**
  - What sort of fire?
  - Where?

- **Assessment:**
  - Relevant from irrelevant
  - Useful from useless

- **Action:**
  - Maybe send someone
  - Evacuate the neighborhood
There's a Fire!

- How to tell the severity of an (incoming) emergency call?

- What is it sort of emergency (fire alarm, first aid call, shooting)?

- What is its context? E.g., for fires
  - is it near a gas station or a pipeline?
  - are there any schools/kindergartens nearby?
  - is a hospital affected?

- What is its context? E.g., for shootings
  - what are possible escape routes?
  - are there any schools/kindergartens nearby?
There's a Fire!

• For answering these questions, background knowledge is required

• E.g., Linked Geo Data
  – Information about objects with coordinates
  – Queries such as: give me all objects within 50m of (lat,long)

• As for incidents, relevance of Linked Geo Data needs to be assessed
  – gas stations and pipelines may be relevant
  – phone booths and statues probably are not
  – based on user-defined rules

www.linkedgeodata.org
Interlude: Linked Geo Data

- Wraps data from Open Street Maps as LOD
- Objects with coordinates
There's a Fire!

• MICI: Live emergency calls from the city of Seattle
  – provided as RSS
• Plus
  – Example Rules
  – Linked Geo Data
There's a Fire!
There's a Fire!

### MICI - Mashup for Identifying Critical Infrastructure

#### Index

Create New

<table>
<thead>
<tr>
<th>Incident Type</th>
<th>Radius</th>
<th>Object Types</th>
<th>Risk Level</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aid Response</td>
<td>50</td>
<td><a href="http://linkedgedata.org/ontology/Parking;fuel">http://linkedgedata.org/ontology/Parking;fuel</a>&lt;br&gt;<a href="http://linkedgedata.org/ontology/Parking;restaurant;fuel">http://linkedgedata.org/ontology/Parking;restaurant;fuel</a>&lt;br&gt;<a href="http://linkedgedata.org/ontology/Fuel;carWash">http://linkedgedata.org/ontology/Fuel;carWash</a>&lt;br&gt;1 more object types</td>
<td>Severe Risk</td>
<td>Edit</td>
</tr>
<tr>
<td>Aid Response</td>
<td>100</td>
<td><a href="http://linkedgedata.org/ontology/OilPlatform">http://linkedgedata.org/ontology/OilPlatform</a>&lt;br&gt;<a href="http://linkedgedata.org/ontology/Pump">http://linkedgedata.org/ontology/Pump</a>&lt;br&gt;<a href="http://linkedgedata.org/ontology/Pipeline">http://linkedgedata.org/ontology/Pipeline</a>&lt;br&gt;2 more object types</td>
<td>Severe Risk</td>
<td>Edit</td>
</tr>
</tbody>
</table>
There's a Fire!

- MICI live: http://mici.tk.informatik.tu-darmstadt.de/
- Recap:
  - Plenty of data (incoming 911 messages)
  - Massive background information (Linked Geo Data)
  - Filtering based on rules
  - Helps: assessing information and acting properly
- Limitations:
  - Data is already preprocessed (RSS from data.seattle.gov)
  - Rules are created manually
Building Knowledge-Intensive Applications with Linked Open Data*

*using examples from the domains sex, drugs, and crime

Chris Bizer, Heiko Paulheim, University of Mannheim
Brief Interlude: Machine Learning Basics

• An essential ingredient to intelligent applications
  – Dealing with new pieces of knowledge
  – Handling unknown situations
  – Adapting to users' needs
  – Making predictions for the future

• Inductive vs. Deductive Reasoning:
  – Deductive: rules + facts $\rightarrow$ facts
  – Inductive: facts $\rightarrow$ rules
Brief Interlude: Machine Learning Basics

- Example: learning a new concept, e.g., "Tree"

- "tree"

- "tree"

- "tree"

- "not a tree"

- "not a tree"

- "not a tree"

- "not a tree"
Brief Interlude: Machine Learning Basics

• Example: learning a new concept, e.g., "Tree"
  – we look at (positive and negative) examples
  – ...and derive a model
    • e.g., "Trees are big, green plants"

• Goal: Classification of new instances

"tree?"

Warning: Models are only approximating examples! Not guaranteed to be correct or complete!
Brief Interlude: Machine Learning Basics

• Typical tasks:
  – Classification (binary or multi-label)
  – Regression (i.e., predicting numerical values)
  – Clustering (finding groups of objects)
  – Frequent Pattern Mining
  – ...

• Methods:
  – Statistical approaches (Naive Bayes, Support Vector Machines, …)
  – Symbolic approaches (Rules, Decision Trees, …)
  – …
Linked Open Data for Machine Learning

- Example machine learning task: predicting book sales

<table>
<thead>
<tr>
<th>ISBN</th>
<th>City</th>
<th>Population</th>
<th>Genre</th>
<th>Publisher</th>
<th>Sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-2347-3427-1</td>
<td>Darmstadt</td>
<td>144402</td>
<td>Crime</td>
<td>Bloody Books</td>
<td>124</td>
</tr>
<tr>
<td>3-43784-324-2</td>
<td>Mannheim</td>
<td>291458</td>
<td>Crime</td>
<td>Guns Ltd.</td>
<td>493</td>
</tr>
<tr>
<td>3-145-34587-0</td>
<td>Roßdorf</td>
<td>12019</td>
<td>Travel</td>
<td>Up&amp;Away</td>
<td>14</td>
</tr>
</tbody>
</table>

→ Crime novels sell better in larger cities

Data Mining Framework "FeGeLOD", RapidMiner Plugins
The FeGeLOD Framework

<table>
<thead>
<tr>
<th>ISBN</th>
<th>City</th>
<th># sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-2347-3427-1</td>
<td>Darmstadt</td>
<td>124</td>
</tr>
</tbody>
</table>

**Named Entity Recognition**

<table>
<thead>
<tr>
<th>ISBN</th>
<th>City</th>
<th>City_URI</th>
<th># sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-2347-3427-1</td>
<td>Darmstadt</td>
<td><a href="http://dbpedia.org/resource/Darmstadt">http://dbpedia.org/resource/Darmstadt</a></td>
<td>124</td>
</tr>
</tbody>
</table>

**Feature Generation**

<table>
<thead>
<tr>
<th>ISBN</th>
<th>City</th>
<th>City_URI</th>
<th>City_URI_dbpedia-owl:populationTotal</th>
<th>City_URI...</th>
<th># sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-2347-3427-1</td>
<td>Darmstadt</td>
<td><a href="http://dbpedia.org/resource/Darmstadt">http://dbpedia.org/resource/Darmstadt</a></td>
<td></td>
<td></td>
<td>141471</td>
</tr>
</tbody>
</table>

**Feature Selection**

<table>
<thead>
<tr>
<th>ISBN</th>
<th>City</th>
<th>City_URI</th>
<th>City_URI_dbpedia-owl:populationTotal</th>
<th># sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-2347-3427-1</td>
<td>Darmstadt</td>
<td><a href="http://dbpedia.org/resource/Darmstadt">http://dbpedia.org/resource/Darmstadt</a></td>
<td></td>
<td>141471</td>
</tr>
</tbody>
</table>
The FeGeLOD Framework

• Entity Recognition
  – Simple approach: guess DBpedia URIs
  – Hit rate >95% for cities and countries (by English name)

• Feature Generation
  – Different Generators
    • Data values (including heuristic numerical conversion)
    • Classes (plus transitive closure)
    • Quantifying Unqualified relations (boolean or numeric)
    • Quantifying Qualified relations (boolean or numeric)

• Feature Selection
  – Filter noise: >95% unknown, identical, or different nominals
The FeGeLOD Prototype (now: RapidMiner Linked Open Data Extension)
The FeGeLOD Prototype
(now: RapidMiner Linked Open Data Extension)
The FeGeLOD Prototype
(now: RapidMiner Linked Open Data Extension)
• So far, we were using preprocessed RSS data
  – How can we detect incidents automatically?
• And handcrafted rules
  – How can we acquire rules?
Detecting Incidents from Social Media

• Social media contains data on many incidents
  – But keyword search is not enough
  – Detecting small incidents is hard
  – Manual inspection is too expensive (and slow)

• Machine learning could help
  – Train a model to classify incident/non incident tweets
  – Apply model for detecting incident related tweets

• Training data:
  – Traffic accidents
  – ~2,000 tweets containing relevant keywords (“car”, “crash”, etc.), hand labeled (50% related to traffic incidents)
Detecting Incidents from Social Media

• Learning to classify tweets:
  – Positive and negative examples
  – Features:
    • Stemming
    • POS tagging
    • Word n-grams
    • …

• Accuracy ~90%

• But
  – Accuracy drops to ~85% when applying the model to a different city
Brief Interlude: Model Overfitting

• What happens here?
  – Model is trained on a sample of labeled data
  – Tries to identify the characteristics of that data

• Possible effect:
  – Model is too close to training data
Brief Interlude: Model Overfitting

• Extreme example
  – Predict credit rating

• Possible (useful) model:
  – (job status = employed) && (debts<5000) → rating=positive

<table>
<thead>
<tr>
<th>Name</th>
<th>Net Income</th>
<th>Job status</th>
<th>Debts</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>40000</td>
<td>employed</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>Mary</td>
<td>38000</td>
<td>employed</td>
<td>10000</td>
<td>-</td>
</tr>
<tr>
<td>Stephen</td>
<td>21000</td>
<td>self-employed</td>
<td>20000</td>
<td>-</td>
</tr>
<tr>
<td>Eric</td>
<td>2000</td>
<td>student</td>
<td>10000</td>
<td>-</td>
</tr>
<tr>
<td>Alice</td>
<td>35000</td>
<td>employed</td>
<td>4000</td>
<td>+</td>
</tr>
</tbody>
</table>
Brief Interlude: Model Overfitting

• Extreme example
  – Predict credit rating

• Possible overfit models:
  – (34000<income<36000) || (income>39000) → rating=positive
  – (name=John) || (name=Alice) → rating=positive

<table>
<thead>
<tr>
<th>Name</th>
<th>Net Income</th>
<th>Job status</th>
<th>Debts</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>40000</td>
<td>employed</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>Mary</td>
<td>38000</td>
<td>employed</td>
<td>10000</td>
<td>-</td>
</tr>
<tr>
<td>Stephen</td>
<td>21000</td>
<td>self-employed</td>
<td>20000</td>
<td>-</td>
</tr>
<tr>
<td>Eric</td>
<td>2000</td>
<td>student</td>
<td>10000</td>
<td>-</td>
</tr>
<tr>
<td>Alice</td>
<td>35000</td>
<td>employed</td>
<td>4000</td>
<td>+</td>
</tr>
</tbody>
</table>
Brief Interlude: Model Overfitting

• All three models perfectly describe the training data
  – But only one is a useful generalization

• Two goals of machine learning (sometimes contradicting):
  – Explain training data as good as possible
  – Find a model that is as general as possible

• Strategies for preventing overfitting:
  – Cross validation
  – Model pruning
  – Stopping criteria in model building
  – Occam's Razor
  – ...
Detecting Incidents from Social Media

• Accuracy ~90%
• But
  – Accuracy drops to ~85% when applying the model to a different city
  – Model overfitting?

• Example set:
  – “Again crash on I90”
  – “Accident on I90”

• Model:
  – “I90” → indicates traffic accident

• Applying the model:
  – “Two cars crashed on I51” → not related to traffic accident
Using LOD for Preventing Overfitting

• Example set:
  – “Again crash on I90”
  – “Accident on I90”

• Model:
  – dbpedia-owl:Road → indicates traffic accident

• Applying the model:
  – “Two cars crashed on I51” → indicates traffic accident

• Using DBpedia Spotlight + FeGeLOD
  – Accuracy keeps up at 90%
  – Overfitting is avoided
Back to Incidents

• So far, we were using preprocessed RSS data
  – How can we detect incidents automatically?

• And handcrafted rules
  – How can we acquire rules?

• Recap: Inductive vs. Deductive Reasoning:
  – Deductive: rules + facts → facts
  – Inductive: facts → rules
What's that Noise?

- A slightly different scenario:
  - Noise measurements from cities

- Create predictions
  - For the rest of the city
  - For new infrastructure

- Note:
  - No handcrafted rules
  - But a fully automatic prediction
    - Based on example measurements
What's that Noise?

- Additional data comes from
  - Linked Geo Data
  - Open Street Maps (Streets: types, lanes and speed limits)
  - Deutscher Wetterdienst
What's that Noise?

- Results:
  - Machine trained model for noise
  - ~80% accuracy in predicting the noise level (six classes)
    - Mean absolute error only 0.077

- This allows to...
  - generate a whole noise map from a small set of observations
  - play “what if” with hypothetical changes
    - e.g., how do speed limits affect noise levels?
What's that Noise?

• From preprocessed RSS data
  – To automatic detection (e.g., Twitter)

• From hand-crafted rules
  – To automatically induced models
Supporting Information Extraction

- **Task**: Event Extraction from Wikipedia
- **Joint work with GESIS (Cologne)**

![Wikipedia History Timeline]

http://www.vizgr.org/historical-events/timeline/
Supporting Information Extraction

• Source Material:

January
- January 1 – Estonia officially adopts the euro currency and becomes the seventeenth eurozone country.[2]  
- January 4 – Tunisian street vendor Mohamed Bouazizi dies after setting himself on fire a month earlier, sparking anti-government protests in Tunisia and later other Arab nations. These protests become known collectively as the Arab Spring.[20]  
- January 9-10 – Southern Sudan holds a referendum on independence. The Sudanese electorate votes in favor of independence, paving the way for the creation of the new state in July.[90]  
- January 11 – Flooding and mudslides in the Brazilian state of Rio de Janeiro kills 503.[77]  
- January 14 – Arab Spring: The Tunisian government falls after a month of increasingly violent protests; President Zine El Abidine Ben Ali flees to Saudi Arabia after 23 years in power.[92]  
- January 24 – 37 people are killed and more than 180 others wounded in a bombing at Domodedovo International Airport in Moscow, Russia.[143][142]

February
- February 11 – Arab Spring: Egyptian President Hosni Mubarak resigns after widespread protests calling for his departure, leaving control of Egypt in the hands of the military until a general election can be held.[15]  
- February 22 – March 14 – Uncertainty over Libyan output causes crude oil prices to rise 20% over a two-week period following the Arab Spring,[141] causing the 2011 energy crisis.

March
- March 11 – A 9.1-magnitude[15] earthquake and subsequent tsunami hit the east of Japan, killing 15,840 and leaving another 3,925 missing. Tsunami warnings are issued in 50 countries and territories. Emergencies are declared at four nuclear power plants affected by the quake.[13]  
- March 15 – Arab Spring: Hamad bin Isa Al Khalifa, King of Bahrain, declares a three-month state of emergency as troops from the Gulf Co-operation Council are sent to quell the civil unrest.[146]  
- March 17 – Arab Spring and the Libyan civil war: The United Nations Security Council votes 10-0 to create a no-fly zone over Libya in response to allegations of government aggression against civilians.[146]  
- March 19 – Arab Spring and the Libyan civil war: In light of continuing attacks on Libyan rebels by forces in support of leader Muammar Gaddafi,[151] military intervention authorized under UNSCR 1973 begins as French fighter jets make reconnaissance flights over Libya.[21]  

April
- April 11 – Former Iranian President Laurent Gbagbo is arrested in his home in Abidjan by supporters of elected President Alasane Ouattara with support from French forces thereby ending the 2010–2011 Ivory Coast crisis and civil war.[22]  

May
- May 1 – U.S. President Barack Obama announces that Osama bin Laden, the founder and leader of the militant group Al-Qaeda, has been killed during an American military operation in Pakistan.[24]  
- May 16 – The European Union agrees to €78 billion rescue deal for Portugal. The bailout loan will be equally split between the European Financial Stabilisation Mechanism, the European Financial Stability Facility, and the International Monetary Fund.[24]  
- May 26 – Former Bosnian Serb Army commander Ratko Mladić, wanted for genocide, war crimes and crimes against humanity, is arrested in Serbia.[250]  

2011 in other calendars
- 2011 in other calendars
- 2011 in other calendars
- 2011 in other calendars
- 2011 in other calendars
- 2011 in other calendars
- 2011 in other calendars
- 2011 in other calendars
- 2011 in other calendars
Supporting Information Extraction

• Event data is automatically extracted
  – Date
  – Textual Description
  – Links to other entities (place, involved people, ...)

• Classification of events required
  – Politics, Culture, Sports, …
  – e.g., for better querying, filtering, …

• Approach: use Machine Learning for classification!
Supporting Information Extraction

• Positive Examples for class politics:
  – 2011, March 15 - German chancellor Angela Merkel shuts down the seven oldest German nuclear power plants.
  – 2010, June 3 – Christian Wulff is nominated for President of Germany by Angela Merkel.

• Negative Examples for class politics:
  – 2010, July 7 – Spain defeats Germany 1-0 to win its semi-final and for its first time, along with Netherlands make the 2010 FIFA World Cup Final.
  – 2012, February 16 – Roman Lob is selected to represent Germany in the Eurovision Song Contest.
• Positive Examples for class *politics*:
  – 2011, March 15 - German chancellor *Angela Merkel* shuts down the seven oldest German nuclear power plants.
  – 2010, June 3 – Christian Wulff is nominated for President of Germany by *Angela Merkel*.

• Negative Examples for class *politics*:
  – 2010, July 7 – Spain defeats Germany 1-0 to win its semi-final and for its first time, along with Netherlands make the 2010 FIFA World Cup Final.
  – 2012, February 16 – Roman Lob is selected to represent Germany in the Eurovision Song Contest.

• Possible learned model:
  – "*Angela Merkel*" → Politics
Supporting Information Extraction

- Possibly Learned Model:
  - "Angela Merkel" → Politics

- There are some problems with that model

- Missing generality (again: overfitting!)
  - 2012, May 13, Elections in North Rhine-Westphalia – Hannelore Kraft is elected to continue as Minister-President, heading an SPD-Green coalition.

- Large amount of training examples required
  - At least one positive and one negative example per politician
  - but training examples are expensive...
Supporting Information Extraction

• Possibly Learned Model:
  – "Angela Merkel" → Politics

• How can we do better?

• Background knowledge from Linked Open Data
  – 2011, March 15 - German chancellor Angela Merkel [class: Politician] shuts down the seven oldest German nuclear power plants.
  – 2012, May 13, Elections in North Rhine-Westphalia – Hannelore Kraft [class: Politician] is elected to continue as Minister-President, heading an SPD-Green coalition.

• Model learned in that case:
  – "[class: Politician]" → Politics
Supporting Information Extraction

• Model learned in that case:
  – "[class: Politician]" → Politics

• Much more general
  – Can also classify events with politicians not contained in the training set

• Less training examples required
  – A few events with politicians, athletes, singers, ... are enough
Supporting Information Extraction

• Experiments on Wikipedia data
  – >10 categories
  – 1,000 labeled examples as training set
  – Classification accuracy: 80%

• Plus:
  – We have trained a language-independent model!
    • often, models are like "elect*" → Politics
Using the Events

- E.g., for annotating time series graphs
Using the Events

• Annotating a Time Series
  – e.g., the stock market price of Apple_Inc.

• Starting point: a link to DBpedia
  – e.g., dbpedia:Apple_Inc.

• Simple approach: retrieve all the events for that entity
  – Problem: low recall

• Naive “improvement”:
  – Also include events for entities linked to dbpedia:Apple_Inc. in DBpedia
  – e.g.: dbpedia:IPhone
  – Problem: extremely low precision
  – Due to frequent entities such as dbpedia:United_States
Relatedness in DBpedia

• Required: entities that are \textit{closely} related to dbpedia:Apple_Inc.

• Problem:
  – There is no notion of proximity, predicate weights, etc. in DBpedia
  – And in LOD in general

• Possible solution: let humans rank proximity
  – Subjective
  – Scales badly

• Required:
  – Automatic approximation
Relatedness in DBpedia

• Good approximation: Normalized Google Distance
  – How frequently do two terms co-occur in websites?
  – E.g., “Apple” and “iPhone” co-occur quite frequently
  – “Apple” and “United States” co-occur less frequently

• Problem:
  – Search engines are not for free (for machine requests)
  – Pairwise ranking of all DBpedia entities would cost much money

• Solution:
  – Retrieve search engine based rankings for a small sample
  – Approximate rankings by machine learning model
Relatedness in DBpedia

- **DBpedia FindRelated Service**
  - Trained on 10,000 statements labeled with NGD
  - >50 Features: network based, linguistic, dataset specific
  - Fair correlation with NGD
  - Live:  
    http://wifo5-21.informatik.uni-mannheim.de:8080/DBpediaFindRelated/

- **Use in Time Series Application:**
  - Increases recall up to 25%
  - Fair tradeoff with precision (<10%)
  - Hundreds of entities added to the search!
Time Series Application – The Big Picture

Start concept: Apple_Inc. → DBpedia

Events: “Tim Cook arrives in China for talks with government officials ...”

Further Concepts: Apple_Inc. IPhone IPad Steve_Jobs
Intermediate Recap

• What we have seen so far:
  – Linked Open Data as background knowledge in various tasks
  – Combination with Machine Learning for intelligent applications
  – Additional dimensions on the data (proximity measures)
Building Knowledge-Intensive Applications with Linked Open Data*

*using examples from the domains sex, drugs, and crime

Chris Bizer, Heiko Paulheim, University of Mannheim
And now for Something Completely Different

• Who are these men?
Statistical Data

- Statistics are very wide spread
  - Quality of living in cities
  - Corruption by country
  - Fertility rate by country
  - Suicide rate by country
  - Box office revenue of films
  - ...
Statistical Data

• Questions we are often interested in
  – Why does city X have a high/low quality of living?
  – Why is the corruption higher in country A than in country B?
  – Will a new film create a high/low box office revenue?

• i.e., we are looking for
  – explanations
  – forecasts (e.g., extrapolations)
MY HOBBY: EXTRAPOLATING

As you can see, by late next month you'll have over four dozen husbands.
Better get a bulk rate on wedding cake.

http://xkcd.com/605/
Statistical Data

- What statistics typically look like
Statistical Data

• There are powerful tools for finding correlations etc.
  – but many statistics cannot be interpreted directly
  – background knowledge is missing

• So where do we get background knowledge from?
  – with as little efforts as possible
Statistical Data

- What we have

<table>
<thead>
<tr>
<th>Country</th>
<th>Fertility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benin</td>
<td>7.6</td>
</tr>
<tr>
<td>Niger</td>
<td>6.69</td>
</tr>
<tr>
<td>Mali</td>
<td>8.44</td>
</tr>
<tr>
<td>Somalia</td>
<td>6.36</td>
</tr>
<tr>
<td>Burundi</td>
<td>9.16</td>
</tr>
<tr>
<td>Burundi</td>
<td>6.14</td>
</tr>
<tr>
<td>Ethiopia</td>
<td>6.92</td>
</tr>
<tr>
<td>Cambodia</td>
<td>5.98</td>
</tr>
<tr>
<td>Angola</td>
<td>5.97</td>
</tr>
<tr>
<td>Republic of the Congo</td>
<td>5.68</td>
</tr>
<tr>
<td>Mozambique</td>
<td>5.46</td>
</tr>
<tr>
<td>Malawi</td>
<td>5.43</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>5.39</td>
</tr>
<tr>
<td>Benin</td>
<td>5.31</td>
</tr>
<tr>
<td>Democratic Republic of the Congo</td>
<td>5.24</td>
</tr>
<tr>
<td>Liberia</td>
<td>5.13</td>
</tr>
<tr>
<td>Guinea</td>
<td>5.1</td>
</tr>
<tr>
<td>Sao Tome and Principe</td>
<td>5.06</td>
</tr>
<tr>
<td>Chad</td>
<td>5.06</td>
</tr>
<tr>
<td>Madagascar</td>
<td>5.02</td>
</tr>
<tr>
<td>Sierra Leone</td>
<td>4.94</td>
</tr>
<tr>
<td>Equatorial Guinea</td>
<td>4.91</td>
</tr>
<tr>
<td>Paraguay</td>
<td>4.9</td>
</tr>
<tr>
<td>Sudan</td>
<td>4.84</td>
</tr>
<tr>
<td>Senegal</td>
<td>4.78</td>
</tr>
<tr>
<td>Gaza Strip</td>
<td>4.74</td>
</tr>
<tr>
<td>Nigeria</td>
<td>4.73</td>
</tr>
<tr>
<td>Comoros</td>
<td>4.72</td>
</tr>
<tr>
<td>Togo</td>
<td>4.69</td>
</tr>
<tr>
<td>Yemen</td>
<td>4.63</td>
</tr>
<tr>
<td>Central African Republic</td>
<td>4.63</td>
</tr>
<tr>
<td>Gabon</td>
<td>4.59</td>
</tr>
<tr>
<td>Guinea-Bissau</td>
<td>4.51</td>
</tr>
<tr>
<td>Equatorial Guinea</td>
<td>4.48</td>
</tr>
<tr>
<td>Mauritania</td>
<td>4.3</td>
</tr>
</tbody>
</table>
Statistical Data

• What we need
...and we've already seen FeGeLOD

<table>
<thead>
<tr>
<th>ISBN</th>
<th>City</th>
<th># sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-2347-3427-1</td>
<td>Darmstadt</td>
<td>124</td>
</tr>
</tbody>
</table>

**Named Entity Recognition**

<table>
<thead>
<tr>
<th>ISBN</th>
<th>City</th>
<th>City_URI</th>
<th># sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-2347-3427-1</td>
<td>Darmstadt</td>
<td><a href="http://dbpedia.org/resource/Darmstadt">http://dbpedia.org/resource/Darmstadt</a></td>
<td>124</td>
</tr>
</tbody>
</table>

**Feature Generation**

<table>
<thead>
<tr>
<th>ISBN</th>
<th>City</th>
<th>City_URI</th>
<th>City_URI_dbpedia-owl:populationTotal</th>
<th># sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-2347-3427-1</td>
<td>Darmstadt</td>
<td><a href="http://dbpedia.org/resource/Darmstadt">http://dbpedia.org/resource/Darmstadt</a></td>
<td>141471</td>
<td>124</td>
</tr>
</tbody>
</table>

**Feature Selection**

<table>
<thead>
<tr>
<th>ISBN</th>
<th>City</th>
<th>City_URI</th>
<th>City_URI_dbpedia-owl:populationTotal</th>
<th># sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-2347-3427-1</td>
<td>Darmstadt</td>
<td><a href="http://dbpedia.org/resource/Darmstadt">http://dbpedia.org/resource/Darmstadt</a></td>
<td>141471</td>
<td>124</td>
</tr>
</tbody>
</table>
Statistical Data

• Adding background knowledge
  – FeGeLOD framework

• Correlation analysis
  – e.g., Pearson Correlation Coefficient

• Rule learning
  – e.g., Association Rule Mining
  – e.g., Subgroup Discovery

• Further data preprocessing
  – depending on approach
  – e.g., discretization
Prototype Tool: Explain-a-LOD

- Loads a statistics file (e.g., CSV)
- Adds background knowledge
- Presents explanations
Presenting Explanations

- Verbalization with simple patterns
  - e.g., negative correlation between population and quality of living
  - "A city which has a low population has a high quality of living"

- Color coding
  - By correlation coefficient, confidence/support of rules, etc.
Statistical Data: Examples

• Data Set: Mercer Quality of Living
  – Quality of living in 216 cities worldwide
  – norm: NYC=100 (value range 23-109)
  – As of 1999
  – http://across.co.nz/qualityofliving.htm

• LOD data sets used in the examples:
  – DBpedia
  – CIA World Factbook for statistics by country
Statistical Data: Examples

• Examples for low quality cities
  – big hot cities \((\text{junHighC} \geq 27 \text{ and } \text{areaTotalKm} \geq 334)\)
  – cold cities where no music has ever been recorded
    \((\text{recordedIn_in} = \text{false and janHighC} \leq 16)\)
  – latitude \(\leq 24\) and longitude \(\leq 47\)
    • a very accurate rule
    • but what's the interpretation?
Statistical Data: Examples

• Data Set: Transparency International
  – 177 Countries and a corruption perception indicator (between 1 and 10)
  – As of 2010
  – http://www.transparency.org/cpi2010/results
Statistical Data: Examples

• Example rules for countries with low corruption
  – HDI > 78%
    • Human Development Index, calculated from live expectancy, education level, economic performance
  – OECD member states
  – Foundation place of more than nine organizations
  – More than ten mountains
  – More than ten companies with their headquarter in that state, but less than two cargo airlines
Statistical Data: Examples

- Data Set: Burnout rates
  - 16 German DAX companies
  - Absolute and relative numbers
  - As of 2011
Statistical Data: Examples

• Findings for burnout rates
  – Positive correlation between turnover and burnout rates
  – Car manufacturers are less prone to burnout
  – German companies are less prone to burnout than international ones
    • Exception: Frankfurt
Statistical Data: Examples

- Data Set: Antidepressives consumption
  - In European countries
  - Source: OECD
Statistical Data: Examples

• Findings for antidepressives consumption
  – Larger countries have higher consumption
  – Low HDI → high consumption
  – By geography:
    • Nordic countries, countries at the Atlantic: high
    • Mediterranean: medium
    • Alpine countries: low
  – High average age → high consumption
  – High birth rates → high consumption
Building Knowledge-Intensive Applications with Linked Open Data*

*using examples from the domains sex, drugs and crime

Chris Bizer, Heiko Paulheim, University of Mannheim
Statistical Data: Examples

• Data Set: Suicide rates
  – By country
  – OECD states
  – As of 2005
Statistical Data: Examples

• Findings for suicide rates
  – Democrats have lower suicide rates than other forms of government
  – High HDI → low suicide rate
  – High population density → high suicide rate
  – By geography:
    • At the sea → low
    • In the mountains → high
  – High Gini index → low suicide rate
    • High Gini index ↔ unequal distribution of wealth
  – High usage of nuclear power → high suicide rates
Statistical Data: Examples

• Data set: sexual activity
  – Percentage of people having sex weekly
  – By country
  – Survey by Durex 2005-2009
  – http://chartsbin.com/view/uya
Statistical Data: Examples

• Findings on sexual activity
  – By geography:
    • High in Europe, low in Asia
    • Low in Island states
  – By language:
    • English speaking: low
    • French speaking: high
  – Low average age → high activity
  – High GDP per capita → low activity
  – High unemployment rate → high activity
  – High number of ISP providers → low activity
Building Knowledge-Intensive Applications with Linked Open Data*

*using examples from the domains sex, drugs, and crime

Chris Bizer, Heiko Paulheim, University of Mannheim
Try it... but be careful!

• Download from
  http://www.ke.tu-darmstadt.de/resources/explain-a-lod

• including a demo video, papers, etc.

• Pitfalls
  – Open world assumption
  – LOD may be noisy
  – Biases
  – ...

08/03/13  Chris Bizer, Heiko Paulheim  84
Try it... but be careful!

- Download from
  
  http://www.ke.tu-darmstadt.de/resources/explain-a-lod

- including a demo video, papers, etc.

http://xkcd.com/552/
Conclusions

• Many tasks require massive background knowledge
  – Can be acquired from LOD
  – E.g., FeGeLOD framework

• Machine learning is often useful to...
  – make sense using Linked Open Data
  – answer non-trivial questions
  – Add additional knowledge dimensions (e.g., similarity)
Credits

• Daniel Hienert, GESIS (WikiEvents)
• Simon Holthausen, TU Darmstadt (Time Series Analysis)
• Frederik Janssen, TU Darmstadt (NoiseMap)
• Jacob Karolus, TU Darmstadt (NoiseMap)
• Evgeny Mitichkin, Uni Mannheim (FeGeLOD)
• Petar Ristoski, Uni Mannheim (Incident Detection from Twitter)
• Axel Schulz, SAP Research/TU Darmstadt (MICI, NoiseMap, Incident Detection from Twitter)
• Dennis Wegener, GESIS (WikiEvents)
Building Knowledge-Intensive Applications with Linked Open Data*

*using examples from the domains sex, drugs and crime

Chris Bizer, Heiko Paulheim, University of Mannheim