Generating Possible Explanations for Statistics from Linked Open Data

Heiko Paulheim
TU Darmstadt, Knowledge Engineering Group
Motivation

- Who are these men?
Motivation

- 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
  - ...
Motivation

- 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
Motivation

- What statistics typically look like
Motivation

- 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
Motivation

- What we have

<table>
<thead>
<tr>
<th>Country</th>
<th>Fertility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Niger</td>
<td>7.6</td>
</tr>
<tr>
<td>Uganda</td>
<td>6.69</td>
</tr>
<tr>
<td>Mali</td>
<td>6.44</td>
</tr>
<tr>
<td>Somalia</td>
<td>6.36</td>
</tr>
<tr>
<td>Burundi</td>
<td>6.16</td>
</tr>
<tr>
<td>Bolivia</td>
<td>6.14</td>
</tr>
<tr>
<td>Ethiopia</td>
<td>6.02</td>
</tr>
<tr>
<td>Zambia</td>
<td>5.96</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>5.08</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>Democratic Republic of the Congo</td>
<td>4.91</td>
</tr>
<tr>
<td>Rwanda</td>
<td>4.9</td>
</tr>
<tr>
<td>Sudan</td>
<td>4.84</td>
</tr>
<tr>
<td>Senegal</td>
<td>4.76</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>4.50</td>
</tr>
<tr>
<td>Guinea-Bissau</td>
<td>4.51</td>
</tr>
<tr>
<td>Equatorial Guinea</td>
<td>4.40</td>
</tr>
<tr>
<td>Mauritania</td>
<td>4.33</td>
</tr>
</tbody>
</table>
Motivation

- What we need
Possible Sources for Background Knowledge

Linking Open Data cloud diagram,
Creating Background Knowledge from Linked Open Data

Named Entity Recognition

<table>
<thead>
<tr>
<th>City</th>
<th>City_URI</th>
<th>index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vancouver</td>
<td><a href="http://dbpedia.org/resource/Vancouver">http://dbpedia.org/resource/Vancouver</a></td>
<td>106</td>
</tr>
</tbody>
</table>

Feature Generation

<table>
<thead>
<tr>
<th>City</th>
<th>City_URI</th>
<th>City_URI.dbpedia-owl:populationTotal</th>
<th>City_URI_...</th>
<th>index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vancouver</td>
<td><a href="http://dbpedia.org/resource/Vancouver">http://dbpedia.org/resource/Vancouver</a></td>
<td>578041</td>
<td>...</td>
<td>106</td>
</tr>
</tbody>
</table>

Feature Selection
Creating Background Knowledge from Linked Open Data

- Step 1: Entity Recognition
  - Map each object in the statistic to a corresponding URI in Linked Open Data
  - E.g.: "Vienna" ↔ http://dbpedia.org/resource/Vienna

- Prototype Implementation
  - naive guessing of DBpedia URIs
  - works great for countries and cities (>95% accuracy)
  - optional typecheck (e.g.: http://dbpedia.org/ontology/City)
Creating Background Knowledge from Linked Open Data

- **Step 2: Feature Generation**
  - Using identified URIs for producing new columns in the table

- **Strategies**
  - Datatype properties (e.g., population)
    - mixed
  - Types (e.g., EuropeanCapitals)
    - boolean
  - Unqualified relations (e.g., outgoing "headquarterOf" relations)
    - numeric or boolean
  - Qualified relations (e.g., outgoing "headquarterOf" relations to entities of type "CarManufacturer")
    - numeric or boolean
Creating Background Knowledge from Linked Open Data

- **Step 3: Feature Selection**
  - Outcome of step 2: ~50,000 columns for some datasets
  - Too much to handle for some algorithms

- **Naive preprocessing strategy**
  - Discard all columns with more than p% of all values that are
    - missing
    - identical
    - different (except numerical features)
  - usually: p=95 or p=99
Generating Hypotheses

- 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
Presenting Hypotheses

- 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.
Prototype Tool: Explain-a-LOD

Basic dataset information
Number of instances: 177
Number of generated features: 41
Source attribute: country
Target attribute: index

A country of type OECDMemberEconomies has high index
Correlation: 0.6164

A country of type EuropeanUnionMemberStates has high index
Correlation: 0.4227

A country of type LeastDevelopedCountries has low index
Correlation: -0.3904

A country of type EuropeanUnionMemberEconomies has high index
Correlation: 0.3858
Example Hypotheses

- **Data Set 1: Mercer Quality of Living**
  - Quality of living in 216 cities worldwide
  - norm: NYC=100 (value range 23-109)
  - As of 1999

- **Data Set 2: Transparency International**
  - 177 Countries and a corruption perception indicator (between 1 and 10)
  - As of 2010
Example Hypotheses

- Data Set: Mercer Quality of Living
- Examples for low quality cities
  - big hot cities \((\text{junHighC} \geq 27 \text{ and } \text{areaTotalKm} \geq 334)\)
  - cold cities where no record has ever been made
    \((\text{recordedIn}_\text{in} = \text{false} \text{ and } \text{janHighC} \leq 16)\)
  - \(\text{latd} \leq 24 \text{ and } \text{longd} \leq 47\)
    - a very accurate rule
    - but what's the interpretation?
Example Hypotheses
Example Hypotheses

- Transparency International Dataset
- 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
Evaluation

- How good are the hypotheses?
- We have to ask humans!

- Two datasets (Mercer and Transparency)
  - Six feature generation strategies each
  - Two hypothesis generation strategies each
    (simple correlations, rule learner)
  - Using the top 3 hypotheses of each combination
  - Manually verbalized
Evaluation

- Evaluation Setup
  - questionnaire with 37+38 questions
  - each hypothesis had to be ranked between 1 (worst) and 5 (best)
  - 18 voluntary participants
    - students and researchers, mostly CS
    - 15 male, 3 female
    - age 24 to 45
  - 15-20 minutes to fill in the questionnaire
Evaluation

- Results Mercer
Evaluation

- Results Transparency International

![Bar chart showing the results of Transparency International evaluations. The chart compares Correlation and Rule Learning across various categories such as Data values, Type, Unqualified relation (boolean), Unqualified relation (numeric), Qualified relation (boolean), Qualified relation (numeric), and Joint. The y-axis represents the ratings ranging from 1 to 5.]
Evaluation

- Overall rating of hypotheses
Evaluation

- Basic observations
  - type features work well
  - joint feature sets do not work better than individual ones
  - no clear trend if rule learning or correlation analysis works better

- Further interesting findings
  - users' rating not correlated with machine rating (correlation factor, rule confidence)
  - imprecise hypotheses are favoured over precise ones
  - DBpedia (and Wikipedia) biases influence hypotheses
  - users prefer coherent rules
    - e.g., temperature=high and rain=low
    - but not temperature=high and numberOfHeavyMetalRecords=high
Current and Future Work

- Improving entity recognition
  - type guessing instead of manual typechecking
- Other feature generation algorithms
  - E.g., using individuals, such as `dbprop:governmentType dbpedia:FederalRepublic`
  - Useful e.g. for films
- Scalability issues
  - intelligent on the fly feature selection
  - generation of "deeper" features
- More datasets
  - freebase
  - World Fact Book
  - Eurostat
Current and Future Work

- Separate plausible from less plausible explanations
  - e.g.: *intl. calling code between 200 and 252*
  - vs. *Third-world countries*

- Presentation of hypotheses
  - Many predicates are not proper verbs (e.g., *headquarter*)
  - Predicates are used in a wrong way
  - Provide justification (e.g., examples)
Try it!

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

- including a demo video, papers, etc.
...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
  - Biases
  - DBpedia is noisy
  - ...
...but be careful!

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

- including a demo video, papers, etc.

[Comic Image: http://xkcd.com/552/]

http://xkcd.com/552/
Generating Possible Explanations for Statistics from Linked Open Data

Heiko Paulheim
TU Darmstadt, Knowledge Engineering Group