Explaining and predicting abnormal expenses at large scale using knowledge graph based reasoning

Freddy Lecue - Accenture Labs Ireland, INRIA France
Jiewen Wu - A*STAR Institute Singapore

Freddy Lecue, Jiewen Wu: Explaining and predicting abnormal expenses at large scale using knowledge graph based reasoning. J. Web Sem. 44: 89-103 (2017)
Value Proposition (1)

- $1.3 billion on employee expense for FY 2016
  - Hotel is #1 expense with $500 million
- Estimated fraud (abuse / intended): $17 million (740 people)
- Estimated noncompliance (non-smart spending): $35 million
- Estimated non-spend optimization: $80 million
Value Proposition (2)

Spend Optimization

The case of Accenture has demonstrated a potential reduction of 7.8% of the overall travel expenses amount by enforcing learnt contextualized rules for future travel related expense items.
All other data required for contextualization, and integration in enterprise knowledge graph is:

- Expenses items from the client
- Source for explanation

<table>
<thead>
<tr>
<th>Source Type</th>
<th>Data Source</th>
<th>Description</th>
<th>Format</th>
<th>Historic (Year)</th>
<th>Size per day (GBytes)</th>
<th>Data Provider</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anomaly</td>
<td>190,000+ unique travellers in 500+ cities recorded for 2015</td>
<td>Min., max. number of respectively 2,521 and 24,800 expenses per city(^a)</td>
<td>CSV</td>
<td>2015</td>
<td>.93 (complete) .41 (aggregated)</td>
<td>Private</td>
</tr>
<tr>
<td>Potential</td>
<td>Social media events e.g., music event, political event</td>
<td>Planned events with small attendance</td>
<td>JSON format</td>
<td>2011</td>
<td>Approx. 94 events per day (.49 GBytes)</td>
<td>Eventbrite</td>
</tr>
<tr>
<td>Explanation</td>
<td></td>
<td>Planned events with large attendance</td>
<td>XML format</td>
<td>2011</td>
<td>Approx. 198 events per day (.39 GBytes)</td>
<td>Eventful</td>
</tr>
<tr>
<td>Media news event</td>
<td></td>
<td>Events reported in the media worldwide</td>
<td>JSON format</td>
<td>2015</td>
<td>Approx. 198 events per day (.76 GBytes)</td>
<td>EventRegistry</td>
</tr>
<tr>
<td>DBpedia</td>
<td>Structured facts extracted from wikipedia</td>
<td>RDF(^d)</td>
<td>-</td>
<td>-</td>
<td>Approx. 33,000+ resources in use (.23 GBytes)</td>
<td>Wikipedia</td>
</tr>
<tr>
<td>Wikidata</td>
<td>Structured data from Wikimedia projects</td>
<td>RDF(^e)</td>
<td>-</td>
<td>-</td>
<td>Approx. 189,000+ resources in use (.63 GBytes)</td>
<td>Freebase Google inc.</td>
</tr>
<tr>
<td>Accenture Categories</td>
<td>Structured is-A taxonomy of event categories</td>
<td>RDF(^f)</td>
<td>-</td>
<td>-</td>
<td>25 resources in use (.001 GBytes)</td>
<td>Accenture inc.</td>
</tr>
<tr>
<td>Spatial</td>
<td>World Map (listing of type, GPS coordinate)</td>
<td>OSM XML</td>
<td>-</td>
<td>-</td>
<td>666 GBytes</td>
<td>Open StreetMap(^g)</td>
</tr>
</tbody>
</table>

\(^a\) Represents the minimum and maximum number of expenses reported by the client.

\(^d\) RDF (Resource Description Framework) is a data model used to represent structured information in a machine-readable format.

\(^e\) Wikidata is a knowledge graph derived from Wikipedia and other sources.

\(^f\) RDF is a specific type of knowledge graph, often used for representing data in a structured format.

\(^g\) Open StreetMap is a service that provides location data and maps.
Underlying Technology: **ML & KRR**

Knowledge Graph Reasoning for Explanation

**Statistical Learning - Anomaly Detection**
for Abnormal Expenses Identification
Anomaly Detection (1)

1. Identifying expenses semantically similar with $x$

2. Learning Rules

3. Semantic Context (knowledge) generation

4. Anomaly detection

Expenses $X$

$\tilde{X} \subset X$

$Sim_T(x, x_4) > m_t$

$\rho \in \mathcal{T}|_{\tilde{X}}$

$\text{support}(\rho) > m_s$

$\text{confidence}(\rho) > m_c$

Terminology $\mathcal{T}_x$
Anomaly Detection (2)

Example of (output) Rules using Knowledge Graph representation

\[
\text{HighAbnormalExpense}(x) \leftarrow \\
\text{expensed}(e, x) \land \text{inCity}(x, c) \land \text{events}(c, \{d\}, ev) \land (2) \\
(\text{Expense} \sqcap \exists \text{type.Accommodation} \sqcap \text{date.}\{d\}) (3) \\
\exists \text{amount.}(\exists \text{higherThan.90\%_Context Expense})(x) \land (4) \\
(\text{Employee} \sqcap \exists \text{career.Manager}(e)) \land (5) \\
(\text{City} \sqcap (\exists p\_size. (\exists \text{lessThan.1Million}))(c) \land (6) \\
(\text{Event} \sqcap \exists \text{category.}(\{\text{Music, Movie}\}))(ev) (7)
\]
Anomaly Detection (3) - Demo

Abnormal Expenses

(what is an abnormal expense)
Explanation (1)

- 12% of data is annotated
- Explanations are retrieved from similar expenses and contexts

(Sample of) Abnormal Expenses (with Semantic Representation)

Semantics

Sample Explanation Computation

Finite State Machine

Semantic Matching

Probabilistic Explanation

Semantic Explanation

Abnormal Expense to be explained

Explanation Results

Level-1 Explanation Type

Level-2 Explanation sub-Type
Explanation (what type of abnormality)
Prediction (1)
Prediction (2) - Demo

Prediction (what type of abnormality and why)
Explanatory Reasoning


In a Nutshell
Impact

• **Travel and expenses business owner:**
  • Holistic (spatial, temporal, categorical) view of (abnormal) expenses and their causes
  • Results that could be used with suppliers to better negotiate price
  • Automatic policy recommendation (e.g., on when, how spending)

• **Expenses auditor:**
  • Minimizing communication between employees / auditors by pinpointing causes of abnormal expenses.

• **Event planner:**
  • Support for recommending when / where to travel (if flexible time / location)
  • Integration with any travel booking system (in discussion for internal integration with CWT)

• **Business traveler / end-user:**
  • Recommendation of best actions for business travelers e.g., Recommendation of alternative booking systems
Explaining and predicting abnormal expenses at large scale using knowledge graph based reasoning

Detecting abnormalities in data is a well established research problem in the database and AI community. Various techniques have been matured over the last two decades to achieve excellent results. However most methods address the problem from a statistic and pure data-centric angle, which in turn limit any interpretation. We elaborated a web application running live with real data from (i) travel and expenses from Accenture, (ii) external data from third party such as Google Knowledge Graph, DBPedia (relational DataBase version of Wikipedia) and social events from Eventful, for explaining abnormalities.

Our last release interprets and explains abnormally high accommodation price as abnormalities in 1000+ cities using live data from Accenture. It also exposes prediction and recommendation. Exposed as a web application manipulating semantic web technologies, our technology consumes and exposes semantic representation ready to be embedded in any in-house application. The core service is planned to be integrated in product offering to flag and explain abnormal expense claims by employees in top 500 fortune companies.