Maverick: Discovering Exceptional Facts from Knowledge Graphs

Gensheng Zhang, Damian Jimenez, Chengkai Li

SIGMOD, Houston, June 14, 2018
Exceptional Facts

Denzel Washington followed Sidney Poitier as only the second black to win the Best Actor award.

Entity of interest: Denzel Washington

Context: Best Actor award winners

Attributes: Ethnicity

Peculiar value: African American (only two satisfy)
**Exceptional Facts**

Denzel Washington followed Sidney Poitier as only the second black to win the Best Actor award.

| Entity of interest | Denzel Washington | Given an entity $x$
| Context | Best Actor award winners | find | A context |
| Attributes | Ethnicity | A set of attributes (subspace) |
| Peculiar value | African American (only two satisfy) |
## Exceptional Facts

**Denzel Washington** followed Sidney Poitier as only the second black to win the Best Actor award.

<table>
<thead>
<tr>
<th><strong>Entity of interest</strong></th>
<th><strong>Denzel Washington</strong></th>
<th><strong>Given an entity $x$</strong></th>
<th><strong>such that</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Context</strong></td>
<td>Best Actor award winners</td>
<td><strong>find</strong> A context</td>
<td>the context has many entities, including $x$</td>
</tr>
<tr>
<td><strong>Attributes</strong></td>
<td>Ethnicity</td>
<td><strong>A set of attributes</strong> (subspace)</td>
<td></td>
</tr>
<tr>
<td><strong>Peculiar value</strong></td>
<td>African American (only two satisfy)</td>
<td></td>
<td>$x$ bears a peculiar value w.r.t. the subspace (few in the context have the value)</td>
</tr>
</tbody>
</table>
Denzel Washington followed Sidney Poitier as only the second black to win the Best Actor award.

This was Brazil's first own goal in World Cup history.

Hillary Clinton becomes first female presidential nominee.
Applications

Computational Journalism
- Fact-finding
- Fact-checking
  - The first female presidential nominee was Victoria Woodhull, not Hillary Clinton (snopes.com)

Data Cleaning

Recommendation Systems
- Friends, news, and product promotion
Exceptional Facts from Knowledge Graphs

What is exceptional about G1?
Modeling

Attributes: labels of incoming/outgoing edges

Subspace: a subset of attributes

G1. awarded-to = CRO
Modeling

Context: entities sharing some common characteristics
Defined by a pattern-variable pair

Goals scored by Brazilian players
Modeling

**Context:** entities sharing some common characteristics

Defined by a pattern-variable pair

Goals scored by Brazilian players
Modeling

Context: entities sharing some common characteristics
Defined by a pattern-variable pair

Goals scored by Brazilian players
What is exceptional about G1?

Among all the goals scored by BRA players, G1 is the only own goal.
Problem Formulation

Input

- Entity of interest $v_0$
- Exceptionality function $\chi$
- Result size $k$

Output

- Top-$k$ (context, subspace) pairs with regard to $\chi$, in which $v_0$ stands out
Challenges

- Number of attribute subspaces: $O(2^{|A_{v_0}|})$
- Number of patterns (contexts): $\Omega(2^{|V_G|})$
Related Work

Outlier detection
- Maverick finds conditions that make an object stand out, although the object may not necessarily be an outlier.

Outlying aspect mining
Challenges in adopting existing algorithms:
- Many assume a single-table model: a graph can be an extremely large and sparse table
- Conjunctive queries on a single table ≠ pattern queries
- Multiple tables: unclear how to handle joins
- Unclear how to handle set values
Maverick

Knowledge Graph

(C_8, A_2)
(C_3, A_5)
(C_2, A_1)

Top-k (C,A) Pairs

Beam B (w=2)

Pattern Search Tree

Pattern Generator

Context Evaluator

\( \chi(v_0, \{a_3\}, C_1) \{a_3\} \)

Exceptionality Evaluator

Set Enumeration Tree

\( \emptyset \rightarrow \{a_2\} \rightarrow \{a_2, a_3\} \chi(v_0, \{a_2, a_3\}, C_1) \)

\( \{a_1\} \rightarrow \{a_1, a_2\} \rightarrow \{a_1, a_2, a_3\} \)
Exceptionality Function $\chi$

$\chi(\nu, A, C) \in \mathcal{R}$

outlierness ($\chi_o$) [Angiuli2009TODS], one-of-the-few ($\chi_f$) [Wu2012KDD], isolation scores ($\chi_i$) [Liu2008ICDM]

Upper bound function

Theorem 4.2

$\text{upper}_o(\nu, A, C) = \sum_{S \in S_A} (p_S)^2 - \frac{(2p_{v.A} + 1) \times |C| - 2}{|C|^2}$

Theorem 4.3

$\text{upper}_f(\nu, A, C) = |\{u | u \in \overline{C}_v, p^A_{u.A} > 1/|C|\}| / |C|$

Theorem 4.4

$\text{upper}_i(\nu, A, C) = 1 - 2^{-q_{v.A} \sum_{S \in S_A \backslash \{v.A\}} p_S \times \log_2 p_S}$
Pattern Generator (PG)
Match–based Pattern Generation

- Construct Partial Order of Valid Patterns

**Theorem 5.4.** Suppose \( P' \) is a child of \( P \in \mathbb{P} \), i.e., \( (P, P') \in E_{\mathbb{P}} \) and thus \( P' \) is a valid pattern with matches. Given any match \( M' \) to \( P' \), there exists a match \( M \) to \( P \) that is a subgraph of \( M' \), i.e., \( \forall M' \in \mathcal{M}_{P'}, \exists M \in \mathcal{M}_{P} \text{ s.t. } V_M \subseteq V_{M'} \text{ and } E_M \subseteq E_{M'} \).
Datasets and Experiments

**WCGoals**

Created based on FIFA.com
11 node types, 13 edge types
49,078 nodes, 158,114 edges

**Film-Award**

A subgraph of Freebase
95 node types, 117 edge types
5,437,628 nodes, 10,879,448 edges
See you in Rio

VLDB2018 demo

Maverick: A System for Discovering Exceptional Facts from Knowledge Graphs