Story Generation From Knowledge Graphs

Patrick Saad

Referee: Prof. Dr. Benno Stein
Referee: Prof. Dr. Norbert Siegmund

Master Thesis | SoSe19 | Bauhaus-Universität Weimar
The Research Problem

Keyword query to graph query

Query Language (Cypher, SPARQL)

MATCH (a:Author)-[r1:AUTHOR_IN]->(p1:Paper)-[r2:CITED_BY]->(p2:Paper)
WHERE p1.year = 2019
WITH a, r1, p1, r2, p2
RETURN a.name AS author, count(r2) AS total
ORDER BY total DESC

Google Search
google.com

Making search knowledge graphs like searching the web

Intuitive
Coherent Text
Subjective
Maybe Unavailable

Hard
Raw Results
Objective
Available

Story Generation

Document Collection
Provide users with a visual method to formulating queries using **facets**
Faceted search interfaces provides query simplification using facets.

Complex queries are still hard to formulate (Author + Year + "Top")

Filtered results contain implicit insights.
Related Work | Social Network Analysis, Distant Reading

Find relationship patterns, influential entities, outliers

Social Network Analysis | Centrality, Louvain Algorithm, etc..
Wolfram Alpha - wolframalpha.com

Distant Reading | Influential Authors In Literature
Illustration by Joon Mo Kang, Stanford Literary Lab
Related Work | Automated Journalism

Automatically generate stories from data

➔ Natural Language Processing
➔ Natural Language Generation
➔ Story Templates

Problems

➔ News reporting without in-depth analysis
➔ Insights are still implicit (influential entities?)

Facets such as Location, Candidate, or Party

750 000 articles
Story Generation Framework | Use Case

Knowledge Graph Setup

Semantic Scholar Open Research Corpus
45 million papers (Computer Science, Neuroscience, Biomedical)

(1) Select all papers with a specific author A

(2) Recursively get incoming/outgoing citations

549,066 Papers, 8124 Authors and 632 Journals

Subset from our knowledge graph built using Neo4j 2 and Cypher 3
Story Generation Framework | Use Case

Insight Discovery

Construct graph queries that compute social performance and influence metrics

Neo4j’s graph algorithms library
Betweenness Centrality, PageRank, etc..

Total Direct Relationships
Paper Citations, Author Collaborations, etc..

Statistics from facets of directly connected nodes
Total/Min/Max/Avg Author h-index, Paper Citations, etc..

Total Indirect Relationships
Nested Paper Citations, Nested Author Collaborations, etc..

Discovering insights from social relationships

1 https://neo4j.com/developer/graph-algorithms
Story Generation Framework | Use Case

Story Generation

Automatically generate stories to communicate the insights

Story Types
4 different story types based on the available facets

Story Templates
2 templates

Story Content
Introduction
Data overview using statistics
Top performing entities
Plot graphs

Total stories by story type for different entity types

<table>
<thead>
<tr>
<th>Story Types</th>
<th>Paper</th>
<th>Author</th>
<th>Journal</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerical facet analysis</td>
<td>8</td>
<td>5</td>
<td>9</td>
<td>22</td>
</tr>
<tr>
<td>Time-filtered numerical facet analysis</td>
<td>448</td>
<td>0</td>
<td>0</td>
<td>488</td>
</tr>
<tr>
<td>Numerical facet correlation analysis</td>
<td>28</td>
<td>10</td>
<td>36</td>
<td>74</td>
</tr>
<tr>
<td>Weaver performance analysis</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>485</td>
<td>16</td>
<td>46</td>
<td>547</td>
</tr>
</tbody>
</table>
Weaver User Interface | Search

**Story count: 78**

- **Time series Analysis | 2012** Top Papers (Using The Total Papers Of Authors)
  - Total Nodes: 204
  - Rank: 1, 13, 14, 15, 25, 26

- **Time series Analysis | 2013** Top Papers (Using The Total Papers Of Authors)
  - Total Nodes: 315
  - Rank: 1, 2, 3

- **Time series Analysis | 2014** Top Papers (Using The Total Papers Of Authors)
  - Total Nodes: 307
  - Rank: 1, 10, 11, 12

- **Time series Analysis | 2018** Top Papers (Using The Maximum H-Index Of Authors)
  - Total Nodes: 35
  - Rank: 1, 9

- **Time series Analysis | 2012** Top Papers (Using The Total Collaboration Of Authors)
  - Total Nodes: 264
  - Rank: 1, 9, 10, 11, 17, 18

- **Time series Analysis | 2013** Top Papers (Using The Total Collaboration Of Authors)
  - Total Nodes: 316
  - Rank: 1, 2, 8

- **Time series Analysis | 2014** Top Papers (Using The Total Collaboration Of Authors)
  - Total Nodes: 307
  - Rank: 1, 7, 8, 14

Knowledge Box provides additional graph insights

**Tim Gollub**

#36 of 8124 (Weaver Score of 39136)

**Featured Authors**
- Michael Volek
- Kristof Komlossy
- Maik Anderka
- Johannes Kiesel
- Amd Oberlander

**Featured Papers**
- Improving the Reproducibility of PAN’s Shared Tasks: - Plagiarism Detection, Author Identification, and Author Profiling
- Overview of the 4th International Competition on Plagiarism Detection
- Ousting ivory tower research: towards a web framework for providing experiments as a service
- Recent Trends in Digital Text Forensics and Its Evaluation - Plagiarism Detection, Author Identification, and Author Profiling
## Weaver User Interface | Knowledge Box

<table>
<thead>
<tr>
<th>Featured Journals</th>
<th>Information Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SIGIR Forum</td>
</tr>
<tr>
<td></td>
<td>CoRR</td>
</tr>
<tr>
<td></td>
<td>D-Lib Magazine</td>
</tr>
<tr>
<td></td>
<td>2012 23rd International Workshop on Database and Expert Systems Applications</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>H Index</td>
<td>10 (#21 of 8124)</td>
</tr>
<tr>
<td>Total Author In</td>
<td>23 (#12 of 8124)</td>
</tr>
<tr>
<td>Total Paper Citations</td>
<td>268 (#1076 of 8124)</td>
</tr>
<tr>
<td>Total Collaborations</td>
<td>51 (#85 of 8124)</td>
</tr>
<tr>
<td>TotalNested Collaborations</td>
<td>354 (#295 of 8124)</td>
</tr>
</tbody>
</table>

### Top Connected Entities

- Separate entity ranking for every social metric
Different insights can reveal different kinds of social influence

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Nested Citations</td>
<td>553 (#538 of 4454)</td>
</tr>
<tr>
<td>Total Incoming Citations</td>
<td>55 (#1164 of 4454)</td>
</tr>
<tr>
<td>Year</td>
<td>2012</td>
</tr>
<tr>
<td>Pagerank</td>
<td>0.195 (#3978 of 4454)</td>
</tr>
<tr>
<td>Total Authors</td>
<td>3 (#1527 of 4454)</td>
</tr>
</tbody>
</table>

Community impact from several aspects
Weaver User Interface | Story Templates
[Analysis] [Weaver Performance Index] Top Authors By Their Overall Performance On Weaver!

The following automatically generated story uses the Open Research Corpus dataset.

**Facets**

- Facet Node Author
- Facet attributes Weaver Performance

The Weaver Performance facet is calculated from all the available node ranks from all generated stories. For every facet we computed, we give points for all nodes based on their performance rank for that facet. The points we add are the inverted rank value of the node given the minimum and maximum rank range.

**Example for a facet X**

Minimum rank = 1 (the highest rank)

Maximum rank = 4488 (the lowest rank value is the total number of papers)

If a node n1 has a rank 1, its Weaver Performance score is 4488. The node n2 with a rank of 2 will correspond to a score of 4487, 3 > 4486, and so on.

The lowest ranked node for X will get just 1 point for its Weaver Performance score.

For each node type (e.g. Paper, Author, Journal), we separately aggregate the individual Weaver Performance scores for each available facet to obtain the global Weaver Performance score of nodes.

This score represents the overall performance of the nodes on Weaver.

**Title and Introduction sections**
### Weaver User Interface | Story Templates

#### Data Overview

<table>
<thead>
<tr>
<th>Author Subset</th>
<th>Min</th>
<th>Weaver Performance</th>
<th>Max</th>
<th>Weaver Performance</th>
<th>Mode</th>
<th>Weaver Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>8124</td>
<td>35</td>
<td>1 (0.01%) authors have this value</td>
<td>40230</td>
<td>1 (0.01%) authors have this value</td>
<td>13785</td>
<td>1 (0.01%) authors have this value</td>
</tr>
</tbody>
</table>

- **Average | Weaver Performance**
  - **20409**
  - 4377 (53.88%) authors have a value below this average
  - 3747 (46.12%) authors have a value equal or above this average

---

Data Overview section
# Weaver User Interface | Story Templates

## Top Results

<table>
<thead>
<tr>
<th>Entity Name</th>
<th>Weaver Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christopher D. Manning</td>
<td>40230 Weaver Performance</td>
</tr>
<tr>
<td>Benno Stein</td>
<td>40227 Weaver Performance</td>
</tr>
<tr>
<td>Paolo Rosso</td>
<td>40145 Weaver Performance</td>
</tr>
<tr>
<td>Efstratios Stamatatos</td>
<td>40053 Weaver Performance</td>
</tr>
<tr>
<td>Iryna Gurevych</td>
<td>40009 Weaver Performance</td>
</tr>
<tr>
<td>W. Bruce Croft</td>
<td>39935 Weaver Performance</td>
</tr>
<tr>
<td>Martin Potthast</td>
<td>39906 Weaver Performance</td>
</tr>
<tr>
<td>Rada Mihalcea</td>
<td>39879 Weaver Performance</td>
</tr>
<tr>
<td>Susan T. Dumais</td>
<td>39878 Weaver Performance</td>
</tr>
<tr>
<td>Moshe Koppel</td>
<td>39874 Weaver Performance</td>
</tr>
<tr>
<td>Shlomo Argamon</td>
<td>39745 Weaver Performance</td>
</tr>
<tr>
<td>Daniel Jurafsky</td>
<td>39723 Weaver Performance</td>
</tr>
<tr>
<td>Chris Callison-Burch</td>
<td>39661 Weaver Performance</td>
</tr>
<tr>
<td>ChengXiang Zhai</td>
<td>39651 Weaver Performance</td>
</tr>
<tr>
<td>Qiang Yang</td>
<td>39648 Weaver Performance</td>
</tr>
</tbody>
</table>

Entities ranked by their facet performance

Interconnected Stories, Entities, and Search Results via hyperlinks

Top performing entities section
## Evaluation using CSUQ

5 participants (expert users)

<table>
<thead>
<tr>
<th>Question Category</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Use (questions 1-8)</td>
<td>1.28</td>
<td>0.40</td>
</tr>
<tr>
<td>Information Quality (questions 9-15)</td>
<td>0.72</td>
<td>0.33</td>
</tr>
<tr>
<td>Interface Quality (questions 16-18)</td>
<td>1.07</td>
<td>0.22</td>
</tr>
<tr>
<td>Overall (questions 1 and 19)</td>
<td>1.70</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Story Generation From Knowledge Graphs

Future Work

Bigger knowledge graph using the cluster *(more resources, framework modifications)*

Generate additional insights *(social network analysis, graph theory, etc..)*

Improve story titles and content *(natural language generation, interactive storytelling,)*

Improve the search interface *(keyword query to graph query, iterative usability testing)*

Better search results ranking
Story Generation from Knowledge Graphs

Patrick Saad

Referee: Prof. Dr. Benno Stein
Referee: Prof. Dr. Norbert Siegmund

Master Thesis | SoSe19 | Bauhaus-Universität Weimar