Story Generation From Knowledge Graphs

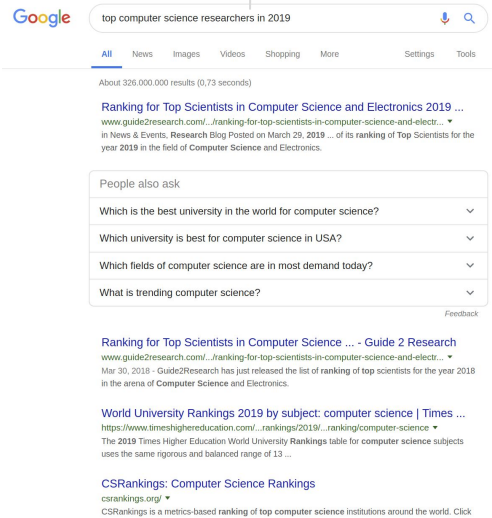
Patrick Saad

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

Master Thesis | SoSe19 | Bauhaus-Universität Weimar

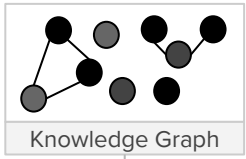
The Research Problem

Document Collection



Google Search
google.com

Keyword query to graph query



Knowledge Graph

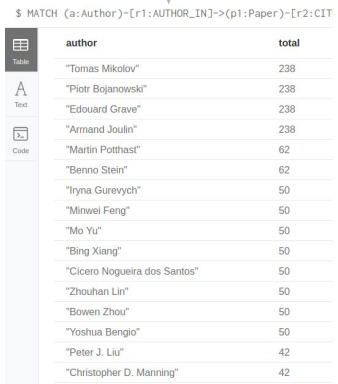
```
MATCH (a:Author)-[r1:AUTHOR_IN]->(p1:Paper)-[r2:CITED_BY]->(p2:Paper)
WHERE p1.year = 2019
WITH a, r1, p1, r2, p2
RETURN a.name AS author, count(r2) AS total
ORDER BY total DESC
```

Query Language (Cypher, SPARQL)

Intuitive
Coherent Text
Subjective
Maybe Unavailable

Story Generation

Hard
Raw Results
Objective
Available



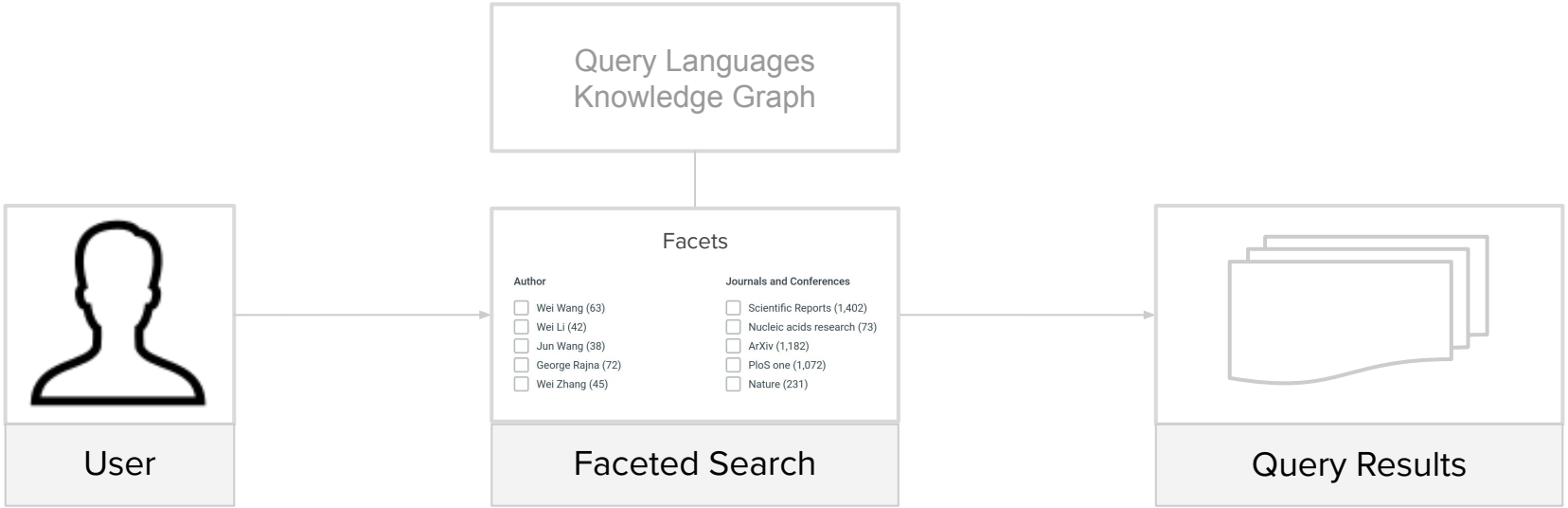
author	total
"Tomas Mikolov"	238
"Piotr Bojanowski"	238
"Edouard Grave"	238
"Armand Joulin"	238
"Martin Potthast"	62
"Benno Stein"	62
"Iryna Gurevych"	50
"Minwei Feng"	50
"Mo Yu"	50
"Bing Xiang"	50
"Cicero Nogueira dos Santos"	50
"Zhouhan Lin"	50
"Bowen Zhou"	50
"Yoshua Bengio"	50
"Peter J. Liu"	42
"Christopher D. Manning"	42

Query Results

Making search knowledge graphs like searching the web

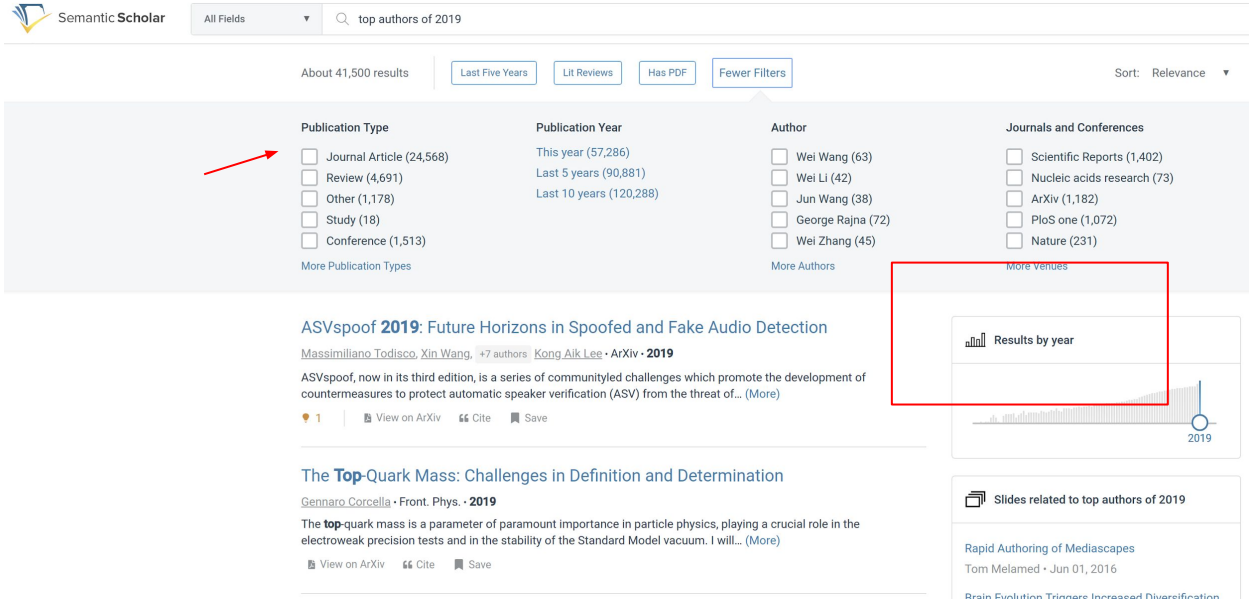
Related Work | Faceted Search Interfaces

Provide users with a visual method to formulating queries using **facets**



Related Work | Faceted Search Interfaces

Faceted search interfaces provides query simplification using **facets**



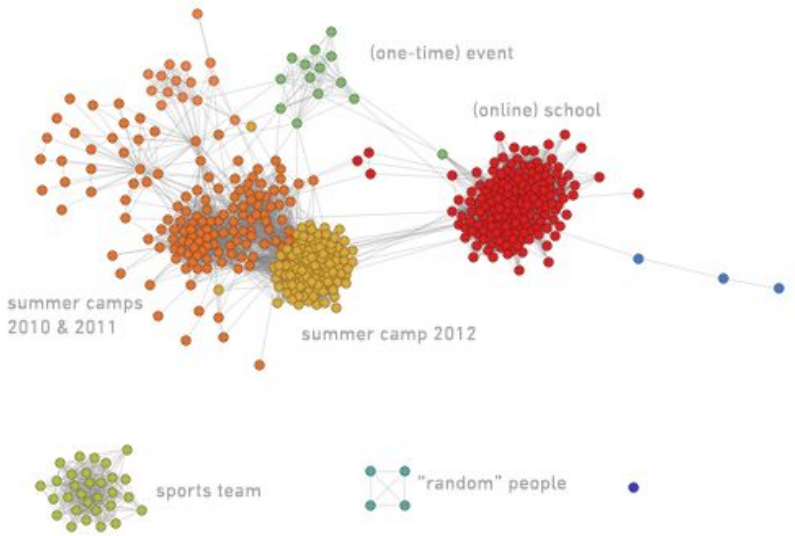
Semantic Scholar
semanticscholar.com

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

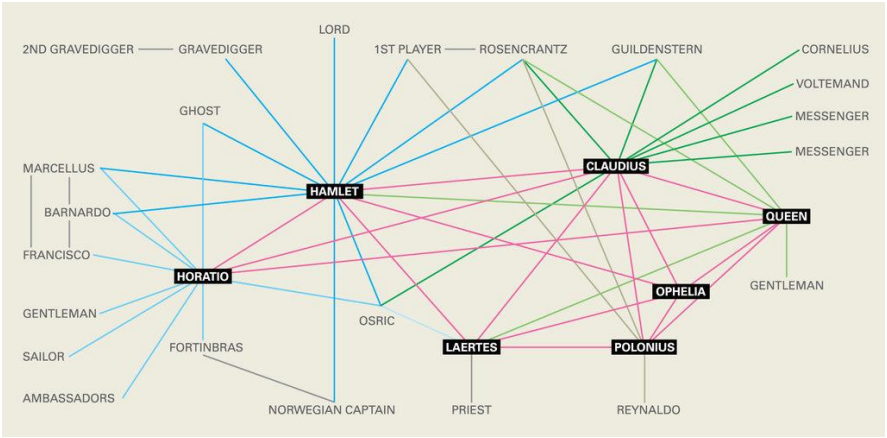
Filtered results contain implicit insights

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

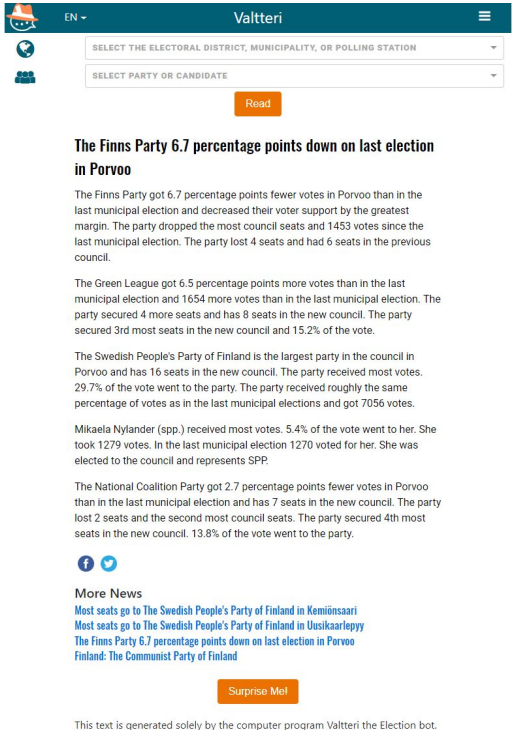
Related Work | Automated Journalism

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?)



Valtteri, the Finnish Municipal Election Bot
vaalibotti.fi

750 000 articles

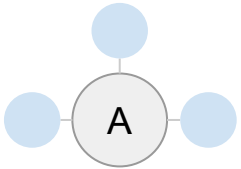
Facets such as Location, Candidate, or Party

Story Generation Framework | Use Case

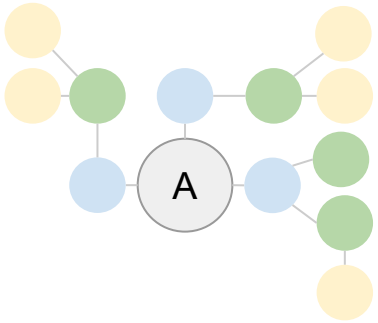
Knowledge Graph Setup

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

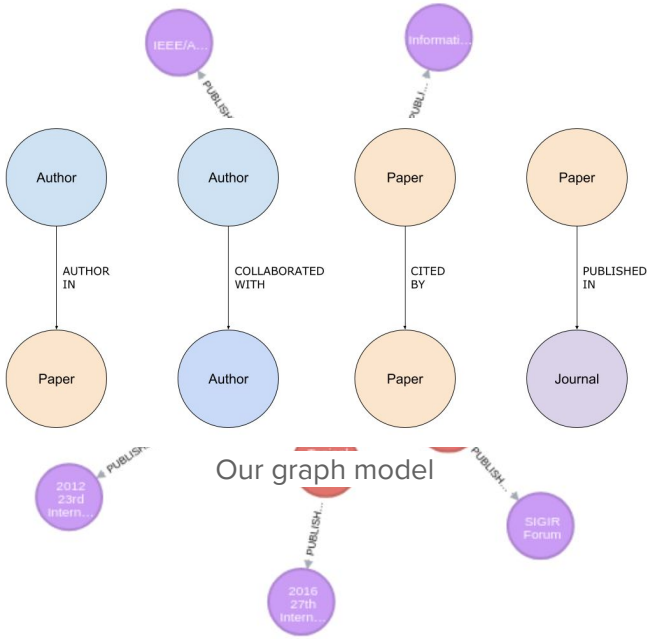
(1) Select all papers with a specific author A



(2) Recursively get incoming/outgoing citations



549,066 Papers, 8124 Authors and 632 Journals



Our graph model

Subset from our knowledge graph built using Neo4j [2] and Cypher [3]

2 Neo4j *https://neo4j.com*
3 Cypher *https://neo4j.com/developer/cypher-query-language*

Story Generation Framework | Use Case

Insight Discovery

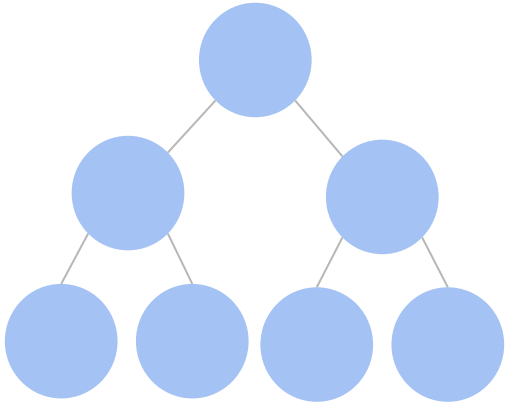
Construct graph queries that compute social performance and influence metrics

Neo4j's graph algorithms library [1]
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

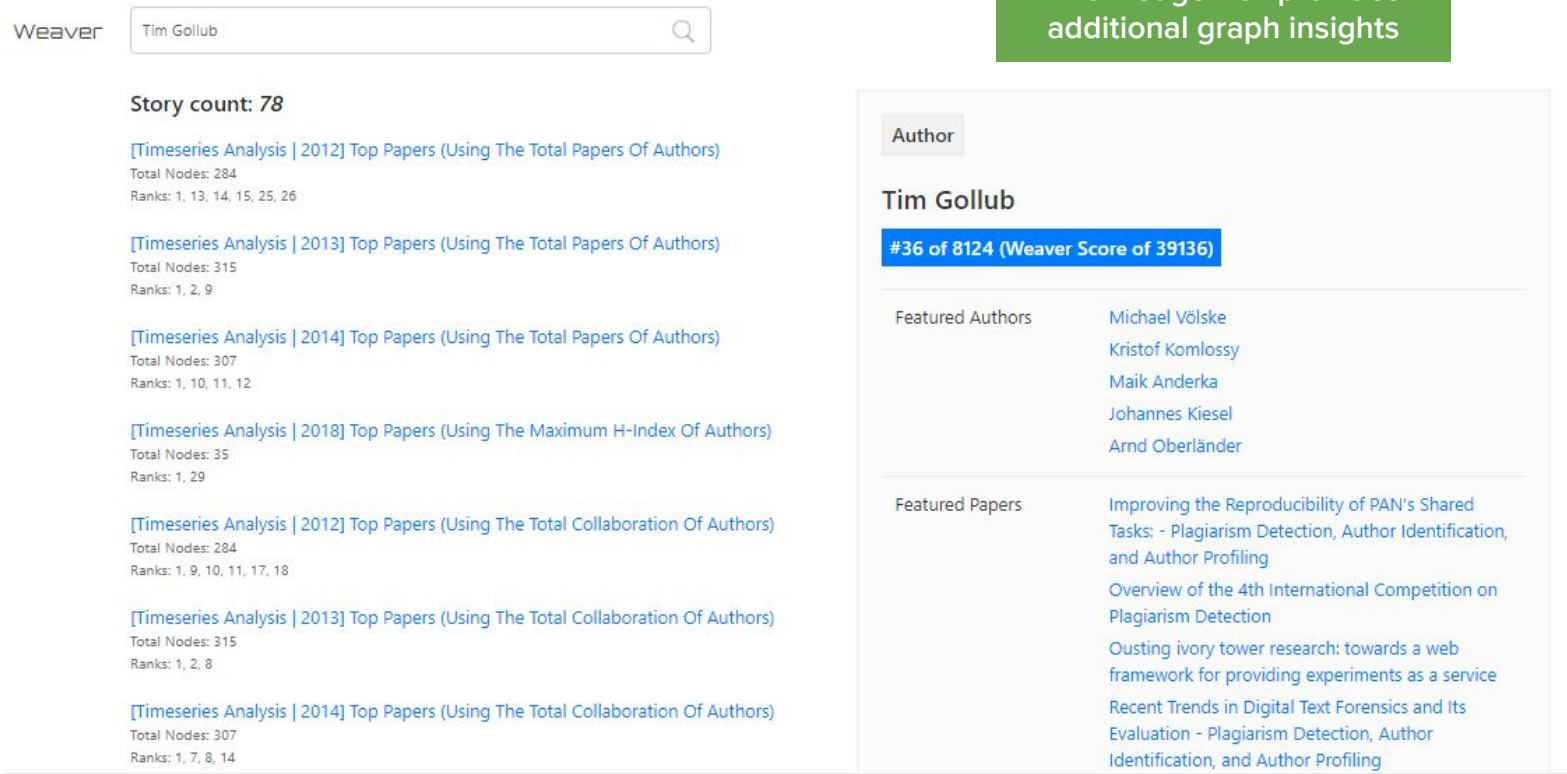
Story Types

	Paper	Author	Journal	Total
Numerical facet analysis	8	5	9	22
Time-filtered numerical facet analysis	448	0	0	488
Numerical facet correlation analysis	28	10	36	74
Weaver performance analysis	1	1	1	3
Total	485	16	46	547

Total stories by story type for different entity types

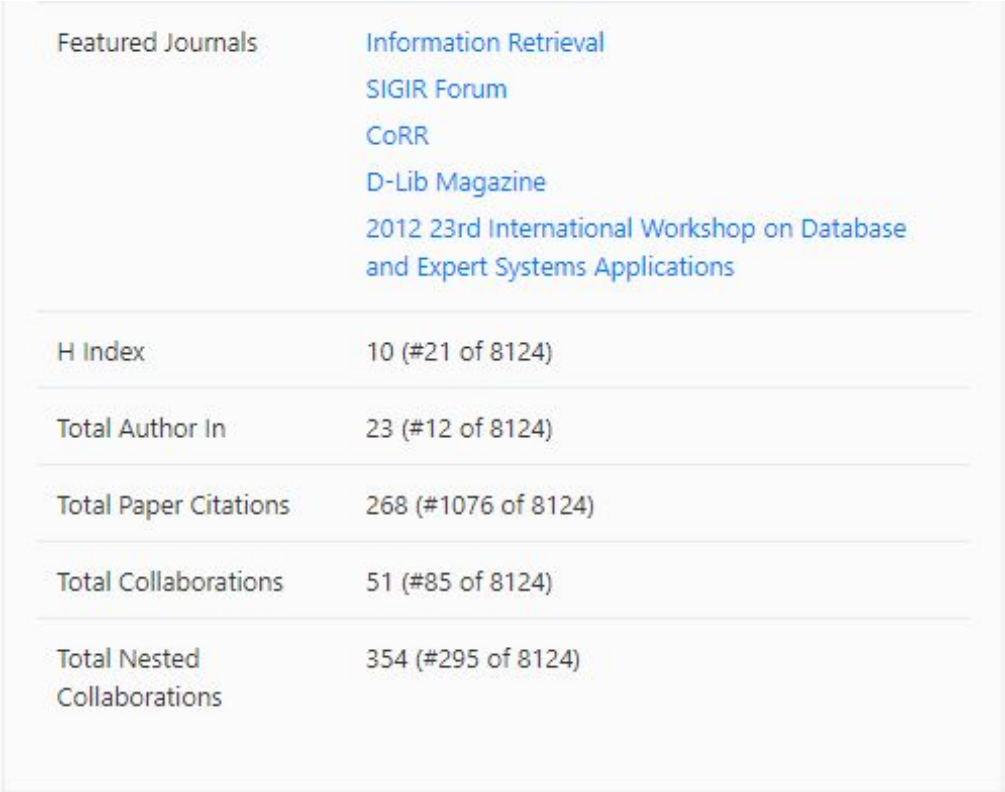
Weaver User Interface | Search

Knowledge Box provides additional graph insights



Example Story Template | Search Results and Knowledge Box

Weaver User Interface | Knowledge Box



Featured Journals	Information Retrieval SIGIR Forum CoRR D-Lib Magazine 2012 23rd International Workshop on Database and Expert Systems Applications
H Index	10 (#21 of 8124)
Total Author In	23 (#12 of 8124)
Total Paper Citations	268 (#1076 of 8124)
Total Collaborations	51 (#85 of 8124)
Total Nested Collaborations	354 (#295 of 8124)

Top Connected Entities

Separate entity ranking for every social metric

Example Story Template | Search Results - Knowledge Box and all facet ranks

Weaver User Interface | Knowledge Box

Different insights can reveal different kinds of social influence

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 attribute: 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.

Data Overview

Author Subset	Min | Weaver Performance	Max | Weaver Performance	Mode | Weaver Performance
8124	35	40230	13785
100.0% of all authors	1 (0.01%) authors have this value	1 (0.01%) authors have this value	1 (0.01%) authors have this value

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

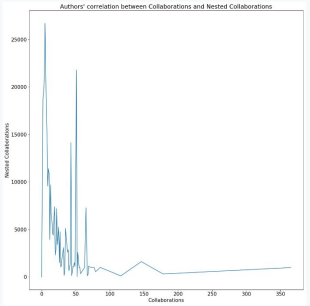
Top Results

Christopher D. Manning 40230 Weaver Performance	Benno Stein 40227 Weaver Performance	Paolo Rosso 40145 Weaver Performance	Efstathios Stamatatos 40053 Weaver Performance	Iryna Gurevych 40009 Weaver Performance
W. Bruce Croft 39835 Weaver Performance	Martin Potthast 39806 Weaver Performance	Rada Mihalcea 39679 Weaver Performance	Susan T. Dumais 39678 Weaver Performance	Moshe Koppel 39674 Weaver Performance
Shlomo Argamon 39745 Weaver Performance	Daniel Jurafsky 39723 Weaver Performance	Chris Callison-Burch 39661 Weaver Performance	ChengXiang Zhai 39651 Weaver Performance	Qiang Yang 39648 Weaver Performance
Eugene Agichtein 39634 Weaver Performance	Philip S. Yu 39593 Weaver Performance	Hinrich Schütze 39554 Weaver Performance	Matthias Hagen 39539 Weaver Performance	Evgeniy Gabrilovich 39524 Weaver Performance

[Correlation Analysis] Authors' Collaborations And Nested Collaborations

Correlation Plot

This correlation analysis uses Pearson's r of the averages of the facets



Negative Correlation

The average Nested Collaborations of a author is decreasing as its Collaborations increases

[Timeseries Analysis | 2017] Top Papers (Using The Total Collaboration Of Authors)

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

Facets

Facet Node: Paper | Facet attribute: Total Collaboration of Authors

Data Overview

Paper Subset	Min | Total Collaboration Of Authors	Max | Total Collaboration Of Authors	Mode | Total Collaboration Of Authors
347	0	227	2
7.79% of all papers	25 (7.2%) papers have this value	1 (0.29%) papers have this value	1 (0.29%) papers have this value

Average | Total Collaboration Of Authors
80
323 (93.08%) papers have a value below this average
24 (6.92%) papers have a value equal or above this average

Top Results

Overview of PAN'17 - Author Identification, Author Profiling, and Author Obfuscation 227.0 Total Collaboration of Authors	Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter 217.0 Total Collaboration of Authors	A Large-Scale Query Spelling Correction Corpus 206.0 Total Collaboration of Authors	Spatio-Temporal Analysis of Reverted Wikipedia Edits 206.0 Total Collaboration of Authors	Overview of the Author Obfuscation Task at PAN 2017: Safety Evaluation Revisited 205.0 Total Collaboration of Authors
Source Retrieval for Web-Scale Text Reuse Detection 205.0 Total Collaboration of Authors	Overview of the Author Identification Task at PAN-2017: Style-Breach Detection and Author Clustering 202.0 Total Collaboration of Authors	TL;DR: Mining Reddit to Learn Automatic Summarization 200.0 Total Collaboration of Authors	An Information Nutritional Label for Online Documents 196.0 Total Collaboration of Authors	Building an Argument Search Engine for the Web 196.0 Total Collaboration of Authors
A Stylometric Inquiry into Hyperpartisan and Fake News 193.0 Total Collaboration of Authors	Overview of the Webis/Isa Vendidikos Detection Task at WSDM Cup 2017 189.0 Total Collaboration of Authors	Patterns of Argumentation Strategies across Topics 186.0 Total Collaboration of Authors	Argumentation Quality Assessment: Theory vs. Practice 178.0 Total Collaboration of Authors	The Argument Reasoning Comprehension Task 175.0 Total Collaboration of Authors
Webis at the CLEF 2017 Dynamic Search Lab 171.0 Total Collaboration of Authors	Unit Segmentation of Argumentative Texts 163.0 Total Collaboration of Authors	WAT-SL: A Customizable Web Annotation Tool for Segment Labeling 162.0 Total Collaboration of Authors	Computational Argumentation Quality Assessment in Natural Language 160.0 Total Collaboration of Authors	The Impact of Modeling Overall Argumentation with Tree Kernels 159.0 Total Collaboration of Authors

weaver.webis.de

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 attribute: Weaver Performance

The Weaver Performance facet is calcualted 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

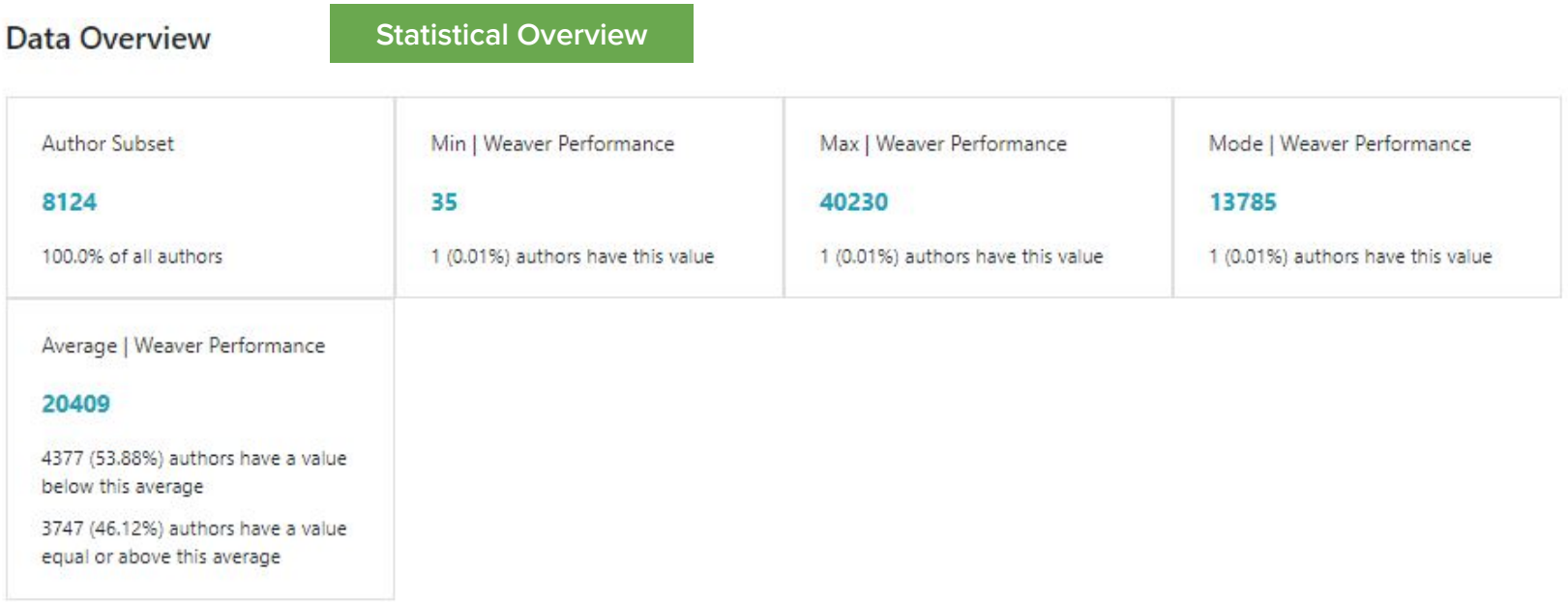
Minimum rank = 1 (the highest rank)
Maximum rank = 4488 (the lowest rank value is the total number of papers)
If a node n1 has a rank 1, its Weaver Performance score is 4488. The node n2 with a rank of 2 will correspond to a score of 4487, 3 > 4486, and so on.
The lowest ranked node for X will get just 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

Introduction (Dataset info, Metric description)

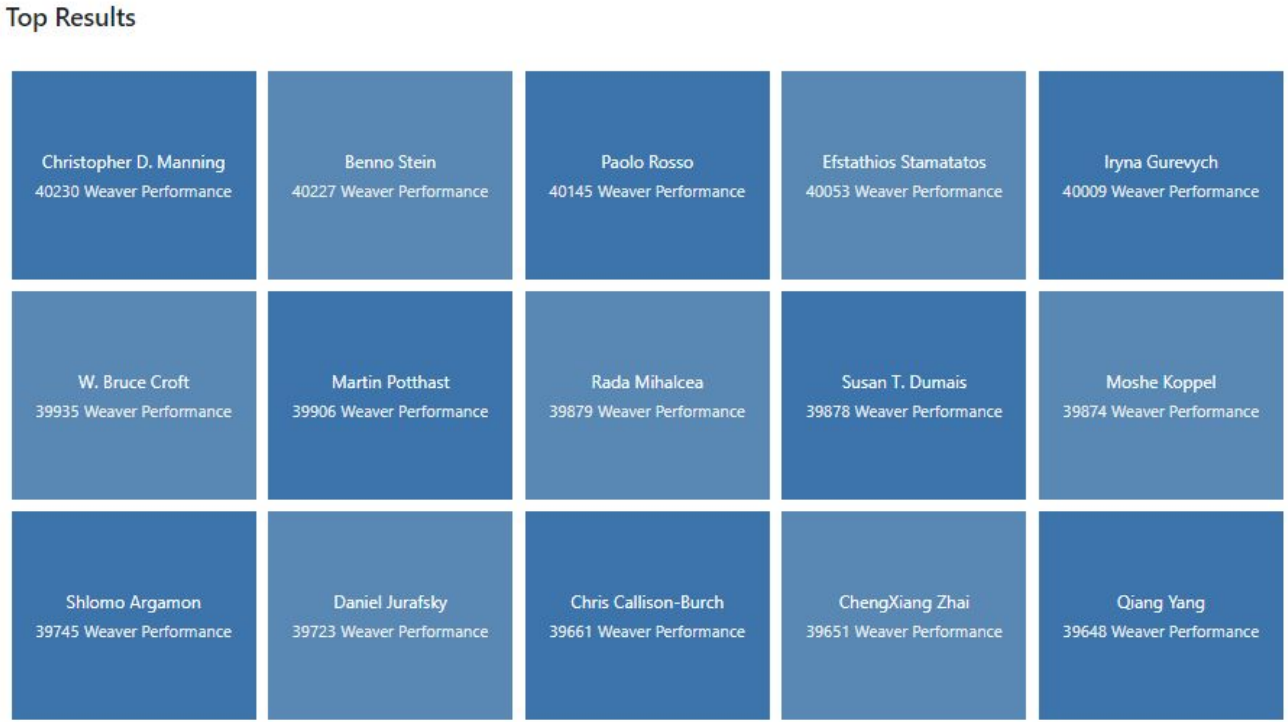
Title and Introduction sections

Weaver User Interface | Story Templates



Data Overview section

Weaver User Interface | Story Templates



Top performing entities section

Entities ranked by their facet performance

Interconnected Stories, Entities, and Search Results via hyperlinks

Story Generation Framework | Use Case

Evaluation using CSUQ

5 participants (expert users)

Strongly disagree — Strongly agree

-3	-2	1	0	1	2	3

Question Category	Mean	Standard Deviation
System Use (questions 1-8)	1.28	0.40
Information Quality (questions 9-15)	0.72	0.33
Interface Quality (questions 16-18)	1.07	0.22
Overall (questions 1 and 19)	**1.70**	0.04

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