

Story Generation From Knowledge Graphs

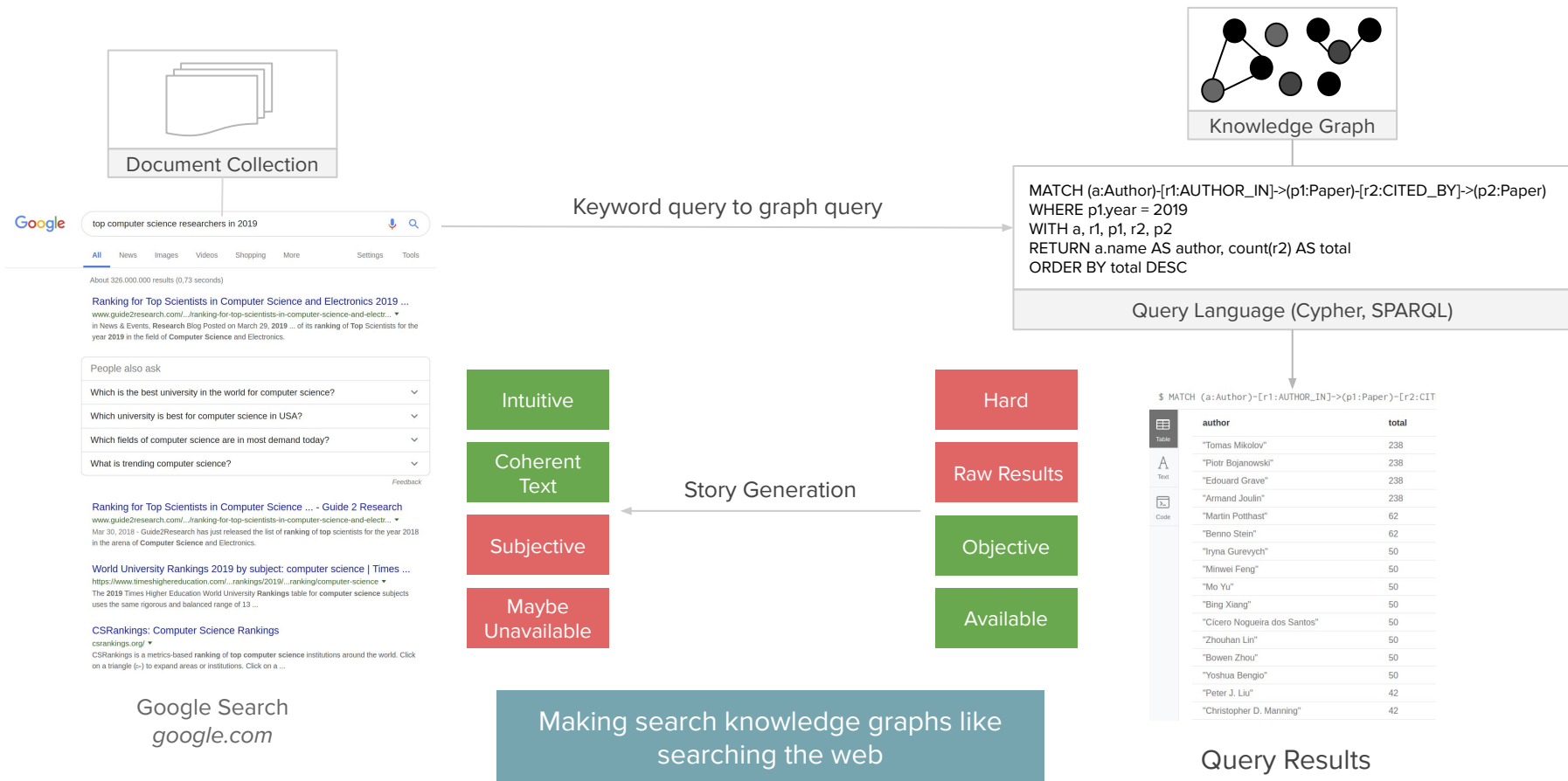
Patrick Saad

Referee: Prof. Dr. Benno Stein

Referee: Prof. Dr. Norbert Siegmund

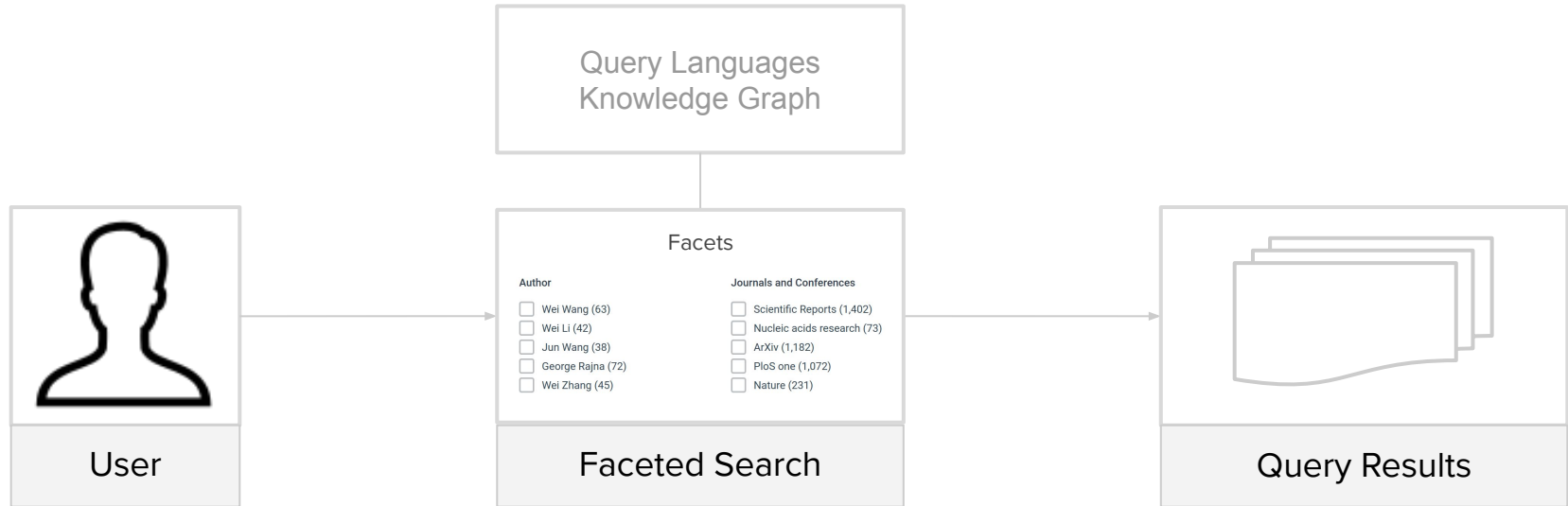
Master Thesis | SoSe19 | Bauhaus-Universität Weimar

The Research Problem



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**

The screenshot displays the Semantic Scholar search results page for the query 'top authors of 2019'. The interface includes a search bar at the top with the query and a dropdown menu for 'All Fields'. Below the search bar, there are filters for 'About 41,500 results', 'Last Five Years', 'Lit Reviews', 'Has PDF', and 'Fewer Filters'. The 'Fewer Filters' button is highlighted with a red arrow. The results are sorted by 'Relevance'. The main content area shows a list of search results, with the first result being 'ASVspooF 2019: Future Horizons in Spoofed and Fake Audio Detection' by Massimiliano Todisco, Xin Wang, and 7 authors, published in ArXiv in 2019. The second result is 'The Top-Quark Mass: Challenges in Definition and Determination' by Gennaro Corcella, published in Front. Phys. in 2019. To the right of the search results, there are four faceted search panels: 'Publication Type' (with options like Journal Article, Review, Other, Study, Conference), 'Publication Year' (with options like This year, Last 5 years, Last 10 years), 'Author' (with options like Wei Wang, Wei Li, Jun Wang, George Rajna, Wei Zhang), and 'Journals and Conferences' (with options like Scientific Reports, Nucleic acids research, ArXiv, PloS one, Nature). A red box highlights the 'Results by year' chart, which shows a bar chart of results over time, with a peak in 2019. Another red box highlights the 'Slides related to top authors of 2019' section, which lists 'Rapid Authoring of Mediascapes' by Tom Melamed, published in June 2016, and 'Brain Evolution Triggers Increased Diversification'.

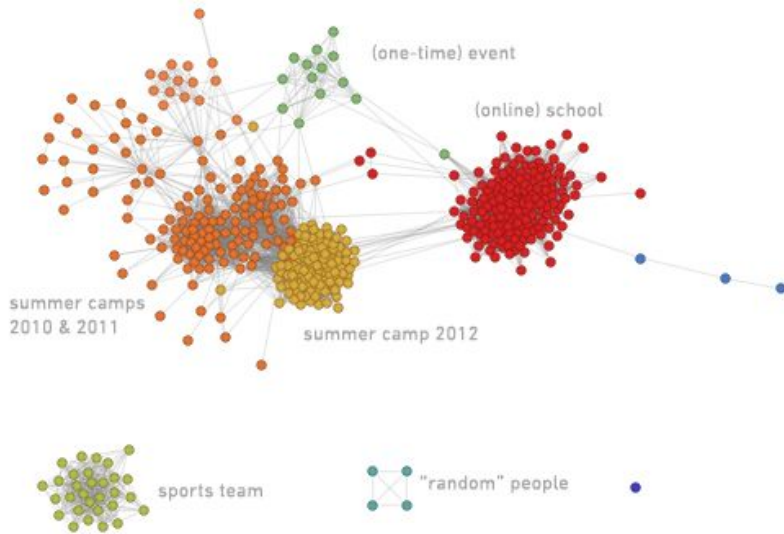
Complex queries are still hard to formulate (Author + Year + "Top")

Filtered results contain implicit insights

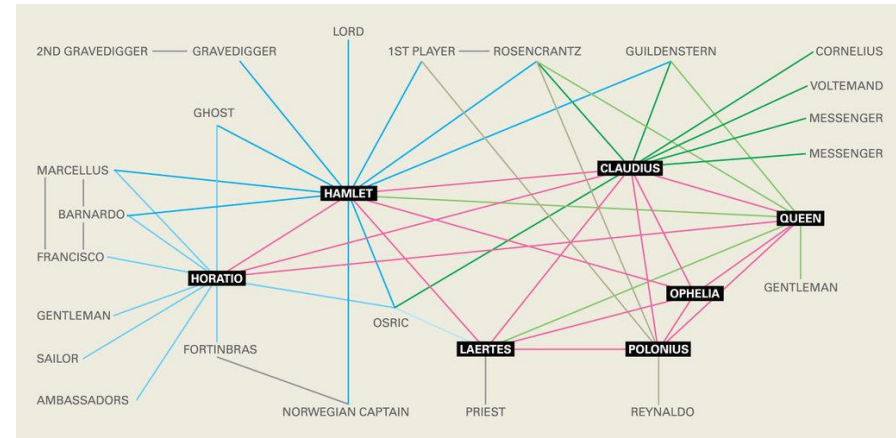
Semantic Scholar
semanticscholar.com

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?)

The screenshot shows the Valtteri web application. At the top, there's a header with a logo, 'EN', and 'Valtteri'. Below the header, there are two dropdown menus: 'SELECT THE ELECTORAL DISTRICT, MUNICIPALITY, OR POLLING STATION' and 'SELECT PARTY OR CANDIDATE'. An orange 'Read' button is positioned below these menus. The main content area displays a news article titled 'The Finns Party 6.7 percentage points down on last election in Porvoo'. The article text describes the party's performance in the last municipal election, noting a decrease in votes and council seats. It also mentions the Green League and the Swedish People's Party of Finland. Below the article, there are social media sharing icons for Facebook and Twitter. A 'More News' section lists related articles with blue links. At the bottom, there is an orange 'Surprise Me!' button and a footer note stating 'This text is generated solely by the computer program Valtteri the Election bot.'

EN Valtteri

SELECT THE ELECTORAL DISTRICT, MUNICIPALITY, OR POLLING STATION

SELECT PARTY OR CANDIDATE

Read

The Finns Party 6.7 percentage points down on last election in Porvoo

The Finns Party got 6.7 percentage points fewer votes in Porvoo than in the last municipal election and decreased their voter support by the greatest margin. The party dropped the most council seats and 1453 votes since the last municipal election. The party lost 4 seats and had 6 seats in the previous council.

The Green League got 6.5 percentage points more votes than in the last municipal election and 1654 more votes than in the last municipal election. The party secured 4 more seats and has 8 seats in the new council. The party secured 3rd most seats in the new council and 15.2% of the vote.

The Swedish People's Party of Finland is the largest party in the council in Porvoo and has 16 seats in the new council. The party received most votes. 29.7% of the vote went to the party. The party received roughly the same percentage of votes as in the last municipal elections and got 7056 votes.

Mikaela Nylander (spp.) received most votes. 5.4% of the vote went to her. She took 1279 votes. In the last municipal election 1270 voted for her. She was elected to the council and represents SPP.

The National Coalition Party got 2.7 percentage points fewer votes in Porvoo than in the last municipal election and has 7 seats in the new council. The party lost 2 seats and the second most council seats. The party secured 4th most seats in the new council. 13.8% of the vote went to the party.

Facebook Twitter

More News

[Most seats go to The Swedish People's Party of Finland in Kemiönsaari](#)

[Most seats go to The Swedish People's Party of Finland in Uusikaarlepyy](#)

[The Finns Party 6.7 percentage points down on last election in Porvoo](#)

[Finland: The Communist Party of Finland](#)

Surprise Me!

This text is generated solely by the computer program Valtteri the Election bot.

750 000 articles

Facets such as
Location, Candidate,
or Party

Valtteri, the Finnish Municipal Election Bot
vaalibotti.fi

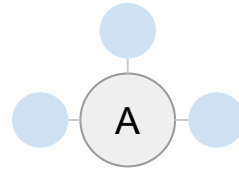
Story Generation Framework | Use Case

Knowledge Graph Setup

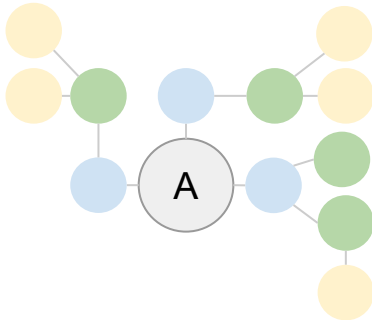
Semantic Scholar Open Research Corpus

45 million papers (Computer Science, Neuroscience, Biomedical)

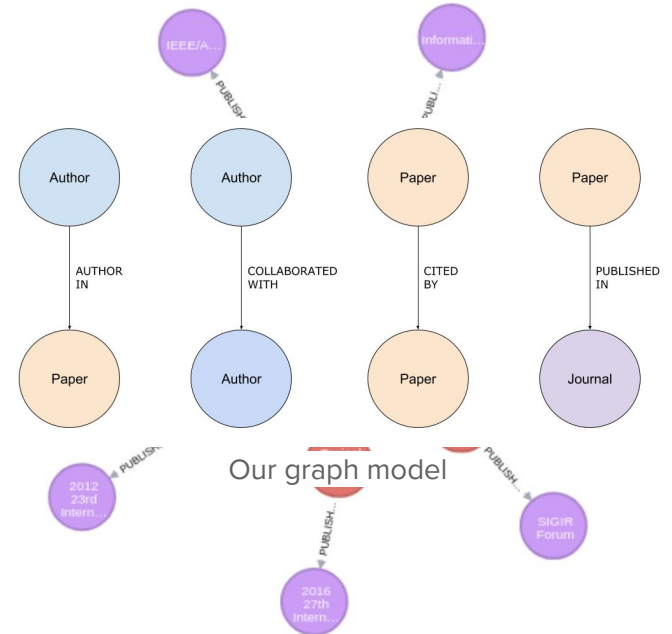
(1) Select all papers with a specific author A



(2) Recursively get incoming/outgoing citations



549,066
Papers, 8124
Authors and
632 Journals



Subset from our knowledge graph built using Neo4j² and Cypher³

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

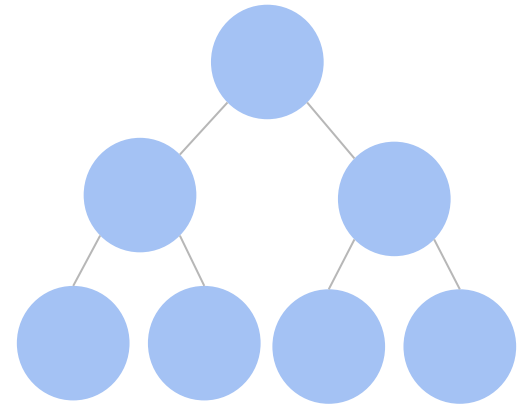
Paper Citations, Author Collaborations, etc..

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

Weaver



Story count: 78

[\[Timeseries Analysis | 2012\] Top Papers \(Using The Total Papers Of Authors\)](#)

Total Nodes: 284

Ranks: 1, 13, 14, 15, 25, 26

[\[Timeseries Analysis | 2013\] Top Papers \(Using The Total Papers Of Authors\)](#)

Total Nodes: 315

Ranks: 1, 2, 9

[\[Timeseries Analysis | 2014\] Top Papers \(Using The Total Papers Of Authors\)](#)

Total Nodes: 307

Ranks: 1, 10, 11, 12

[\[Timeseries Analysis | 2018\] Top Papers \(Using The Maximum H-Index Of Authors\)](#)

Total Nodes: 35

Ranks: 1, 29

[\[Timeseries Analysis | 2012\] Top Papers \(Using The Total Collaboration Of Authors\)](#)

Total Nodes: 284

Ranks: 1, 9, 10, 11, 17, 18

[\[Timeseries Analysis | 2013\] Top Papers \(Using The Total Collaboration Of Authors\)](#)

Total Nodes: 315

Ranks: 1, 2, 8

[\[Timeseries Analysis | 2014\] Top Papers \(Using The Total Collaboration Of Authors\)](#)

Total Nodes: 307

Ranks: 1, 7, 8, 14

Knowledge Box provides
additional graph insights

Author

Tim Gollub

#36 of 8124 (Weaver Score of 39136)

Featured Authors

Michael Völske

Kristof Komlossy

Maik Anderka

Johannes Kiesel

Arnd Oberländer

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](#)

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

Total Nested Citations	553 (#538 of 4454)	Community impact from several aspects
Total Incoming Citations	55 (#1164 of 4454)	
Year	2012	
Pagerank	0.195 (#3978 of 4454)	
Total Authors	3 (#1527 of 4454)	

Different insights can reveal different kinds of social influence

[Analysis] [Weaver Performance Index] Top Authors By Their Overall Performance On Weaver!

Ence

Downloaded from <http://ajphaphysocpharm.com/>

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 n_1 has a rank 1, its Weaver Performance score is 4488. The node n_2 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

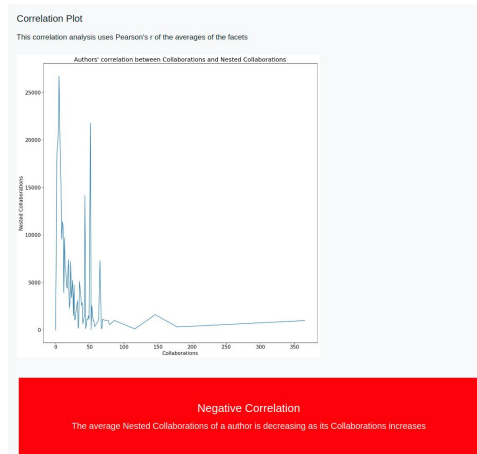
Author Subject	Min Weakest Performance	Max Weakest Performance	Made Weakest Performance
8124	35	40220	13785
100.0% of all authors	1 (0.02%) authors have this value	1 (0.01%) authors have this value	1 (0.02%) authors have this value
Average Weakest Performance			
20409			
4377 (34.88%) authors have a value below this average			
3167 (25.24%) authors have a value equal or above this average			

Top Results

 <p>Christopher D. Manning 4026 Women Performance</p>	 <p>Berna Stein 4027 Women Performance</p>	 <p>Paolo Rosso 4040 Women Performance</p>	 <p>Elizabeth Iremonger 4093 Women Performance</p>	 <p>Iryna Gurych 4000 Women Performance</p>
 <p>W. Brian Croft 3995 Women Performance</p>	 <p>Martin Posthum 3996 Women Performance</p>	 <p>Rade Mihajlovic 3979 Women Performance</p>	 <p>Brian T. Dunais 3913 Women Performance</p>	 <p>Mathé Kippel 3974 Women Performance</p>
 <p>Shano Argenson 3943 Women Performance</p>	 <p>Daniel Josabily 3923 Women Performance</p>	 <p>Chris Carlson Burch 3961 Women Performance</p>	 <p>Chengxiang Zhai 3961 Women Performance</p>	 <p>Qing Yang 3968 Women Performance</p>
 <p>Eugene Agafonov 3954 Women Performance</p>	 <p>Philip S. Yu 3951 Women Performance</p>	 <p>Henrich Schürze 3954 Women Performance</p>	 <p>Mathias Hagim 3939 Women Performance</p>	 <p>Eugeny Gaidarbush 3952 Women Performance</p>



[Correlation Analysis] Authors' Collaborations And Nested Collaborations



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

Facets

Facet Model: Paper Facet attribute: Total Collaboration of Authors

Data Overview

Pages Submit	Min Total Collaboration Of Authors	Max Total Collaboration Of Authors	Mean Total Collaboration Of Authors
347	0	227	2
7.9% of 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 (33.08%) papers have a value below this average			
24 (6.4%) papers have a value equal or above this average			

Top Results

Overview of PNAS 17 Author Identification, Author Profiling, and Author Similarity	Chiffolleau et al. PNAS 2017 Center for Complex Vulnerability Assessment 2017 21/17 Total Collaborations of Authors	A Large-Scale Query Spotting Correlation Coefficients	Spatio-Temporal Analysis of Recurrent Collaborations of Authors	Overview of PNAS 17 Author Identification, Author Profiling, and Author Similarity
2005	2005	2005	2005	2005
Total Collaborations of Authors	Total Collaborations of Authors	Total Collaborations of Authors	Total Collaborations of Authors	Total Collaborations of Authors
Source Retrieval for Bibliometric Search Results	Overview of PNAS 17 Author Identification, Author Profiling, and Author Similarity	17,131 Mining Results to Least Ambiguous Authorship	An Information-Network View for Social Connections	Building an Integrated Search Engine for the Web (Reviewed)
2005	2005	2005	2005	2005
Total Collaborations of Authors	Total Collaborations of Authors	Total Collaborations of Authors	Total Collaborations of Authors	Total Collaborations of Authors
A Systematic Inquiry into Hyperlinking and Author Similarity	Overview of the Wikipedia Vandalism Problem and its Mitigation on Pedia.js	Patterns of Agreement Between Agents Topics	Aggregation+Quality Assessment: Theory vs. Practice	Comparing Researching Conferences
2005	2005	2005	2005	2005
Total Collaborations of Authors	Total Collaborations of Authors	Total Collaborations of Authors	Total Collaborations of Authors	Total Collaborations of Authors
WebVis at the CLEF 2012 System Overview	Web Segmentation of Vandalism	MAE-3.0, a Customizable Web Annotation Tool for Topic Segmentation	Complexity-aware Aggregation+Quality Assessment: The Value of Language	The Impact of Modeling Complexity on Natural Language
2015	2015	2015	2015	2015
Total Collaborations of Authors	Total Collaborations of Authors	Total Collaborations of Authors	Total Collaborations of Authors	Total Collaborations of Authors

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.

Title

Introduction (Dataset
info, Metric
description)

Title and Introduction sections

Weaver User Interface | Story Templates

Data Overview

Statistical Overview

<p>Author Subset</p> <p>8124</p> <p>100.0% of all authors</p>	<p>Min Weaver Performance</p> <p>35</p> <p>1 (0.01%) authors have this value</p>	<p>Max Weaver Performance</p> <p>40230</p> <p>1 (0.01%) authors have this value</p>	<p>Mode Weaver Performance</p> <p>13785</p> <p>1 (0.01%) authors have this value</p>
<p>Average Weaver Performance</p> <p>20409</p> <p>4377 (53.88%) authors have a value below this average</p> <p>3747 (46.12%) authors have a value equal or above this average</p>			

Data Overview section

Weaver User Interface | Story Templates

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 39935 Weaver Performance	Martin Potthast 39906 Weaver Performance	Rada Mihalcea 39879 Weaver Performance	Susan T. Dumais 39878 Weaver Performance	Moshe Koppel 39874 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

Entities ranked by their
facet performance

Interconnected Stories,
Entities, and Search
Results via hyperlinks

Top performing entities section

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