Court Case Page Rank

Time-aware Bibliographic Ranking Algorithm for Court Cases based on Google PageRank Algorithm

Krishna Singh, 3rd year CS and Math major

April 11, 2024

Mathematics Department, Northeastern University College of Science

Table of contents

Introduction

Introduction

This research presents a novel approach to ranking court cases using a time-aware bibliographic PageRank algorithm, adapting Google's famous algorithm to consider factors unique to legal documents.

Motivation

- Traditional bibliographic measures do not account for the temporal aspect of court cases.
- A need for a dynamic ranking system that reflects the evolving nature of legal precedents.
- Enhance legal research efficiency by providing more relevant search and reference tools.

PageRank Algorithm

Google's PageRank Algorithm

- Developed by Larry Page and Sergey Brin in 1996.
- Ranks web pages based on link structures.
- Iteratively transfers rank through links, simulating a "random surfer".

$$PR(A) = (1-d) + d(\frac{PR(T_1)}{C(T_1)} + ... + \frac{PR(T_n)}{C(T_n)})$$

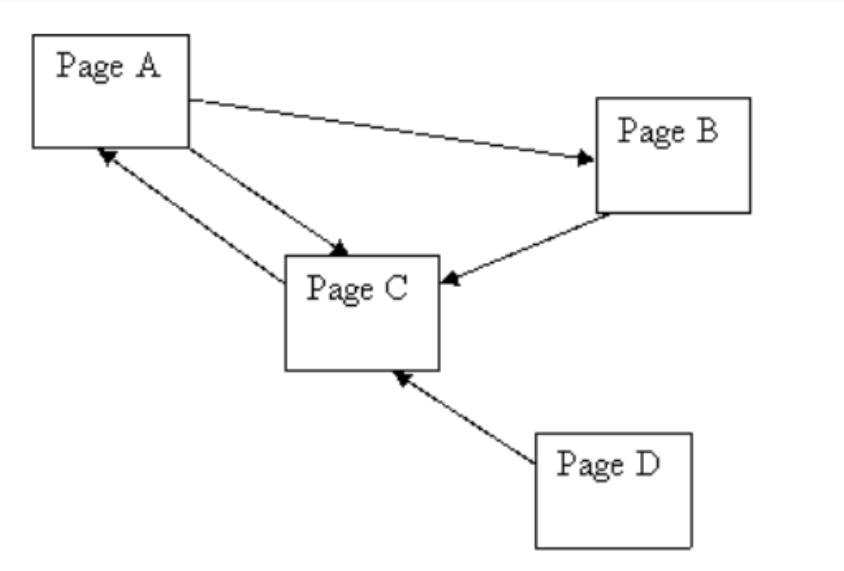
where d is the damping factor

Small Example

Simplified PageRank Calculation Example

Simple Network:

- Two pages, A and B, each linking to the other.
- Each page has one outgoing link, hence $C(A) = 1$ and $C(B) = 1$.



Simplified PageRank Calculation Example

Simple Network:

- Two pages, A and B, each linking to the other.
- Each page has one outgoing link, hence $C(A) = 1$ and $C(B) = 1$.

PageRank Formula:

$$PR(X) = (1 - d) + d \left(\frac{PR(Y)}{C(Y)} \right)$$

where d is the damping factor, typically set to 0.85.

Guess 1: Initial PR of 1.0 for both pages

- $PR(A) = PR(B) = 1$ (No change, lucky guess!)

Simplified PageRank Calculation Example

Simple Network:

- Two pages, A and B, each linking to the other.
- Each page has one outgoing link, hence $C(A) = 1$ and $C(B) = 1$.

PageRank Formula:

$$PR(X) = (1 - d) + d\left(\frac{PR(Y)}{C(Y)}\right)$$

where d is the damping factor, typically set to 0.85.

Guess 2: Initial PR of 0 for both pages

- Iteration 1: $PR(A) = 0.15$, $PR(B) = 0.2775$
- Iteration 2: $PR(A) = 0.385875$, $PR(B) = 0.47799375$
- Numbers keep increasing but never exceed 1.0 due to normalization.

Simplified PageRank Calculation Example

Simple Network:

- Two pages, A and B, each linking to the other.
- Each page has one outgoing link, hence $C(A) = 1$ and $C(B) = 1$.

PageRank Formula:

$$PR(X) = (1 - d) + d \left(\frac{PR(Y)}{C(Y)} \right)$$

where d is the damping factor, typically set to 0.85.

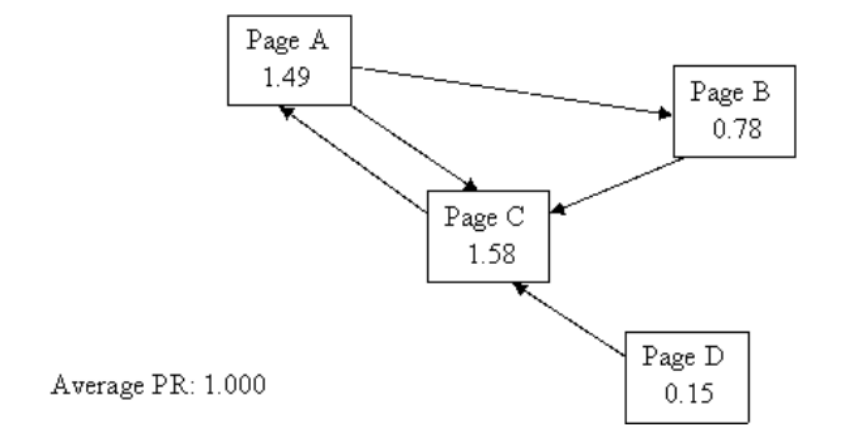
Guess 3: Initial PR of 40 for both pages

- The PageRank values decrease each iteration and approach 1.0.

Simplified PageRank Calculation Example

Simple Network:

- Two pages, A and B, each linking to the other.
- Each page has one outgoing link, hence $C(A) = 1$ and $C(B) = 1$.



Guess 3 is correct!

Simplified PageRank Calculation Example

Simple Network:

- Two pages, A and B, each linking to the other.
- Each page has one outgoing link, hence $C(A) = 1$ and $C(B) = 1$.

PageRank Formula:

$$PR(X) = (1 - d) + d \left(\frac{PR(Y)}{C(Y)} \right)$$

where d is the damping factor, typically set to 0.85.

Principle:

- The final PageRank values converge to a normalized distribution with an average of 1.0.

Efficiency:

- The damping factor and the order of calculations can affect the speed of convergence.
- A simple network may settle after around 20 iterations.

Our Application and Modifications

Adapting PageRank for Court Cases

- Personalization vector adjusted by case date to prioritize newer information.
- Transitions only from newer to older cases, reflecting citation flows.
- Not accounting for geographic and citation relevance yet

Personalization Vector

- A personalization vector is a way to introduce bias or preference into the calculation of PageRank scores, making the algorithm ”personalized” for specific needs.
- The vector typically represents a probability distribution over all nodes in the graph, indicating where a ”random surfer” is more likely to begin or jump to when navigating the network.

Method and Results

How We Obtained the Data

- Data was obtained by web scraping Google Scholar Case Law
- Data ranges from 1940s to 2023
- Visualization done on 500 cases

Web Scraping

- The process involved *Requests* and *Beautiful Soup* in Python for data retrieval and automation.
- After receiving the HTML content, we parsed it using Beautiful Soup to extract articles.
- Each article's title, authors, and URL were collected, along with citations through parsing the json file.
- The script looped over multiple pages to accumulate comprehensive data.

This method allowed us to quickly aggregate structured academic data, significantly accelerating the research process.

Web Scraping

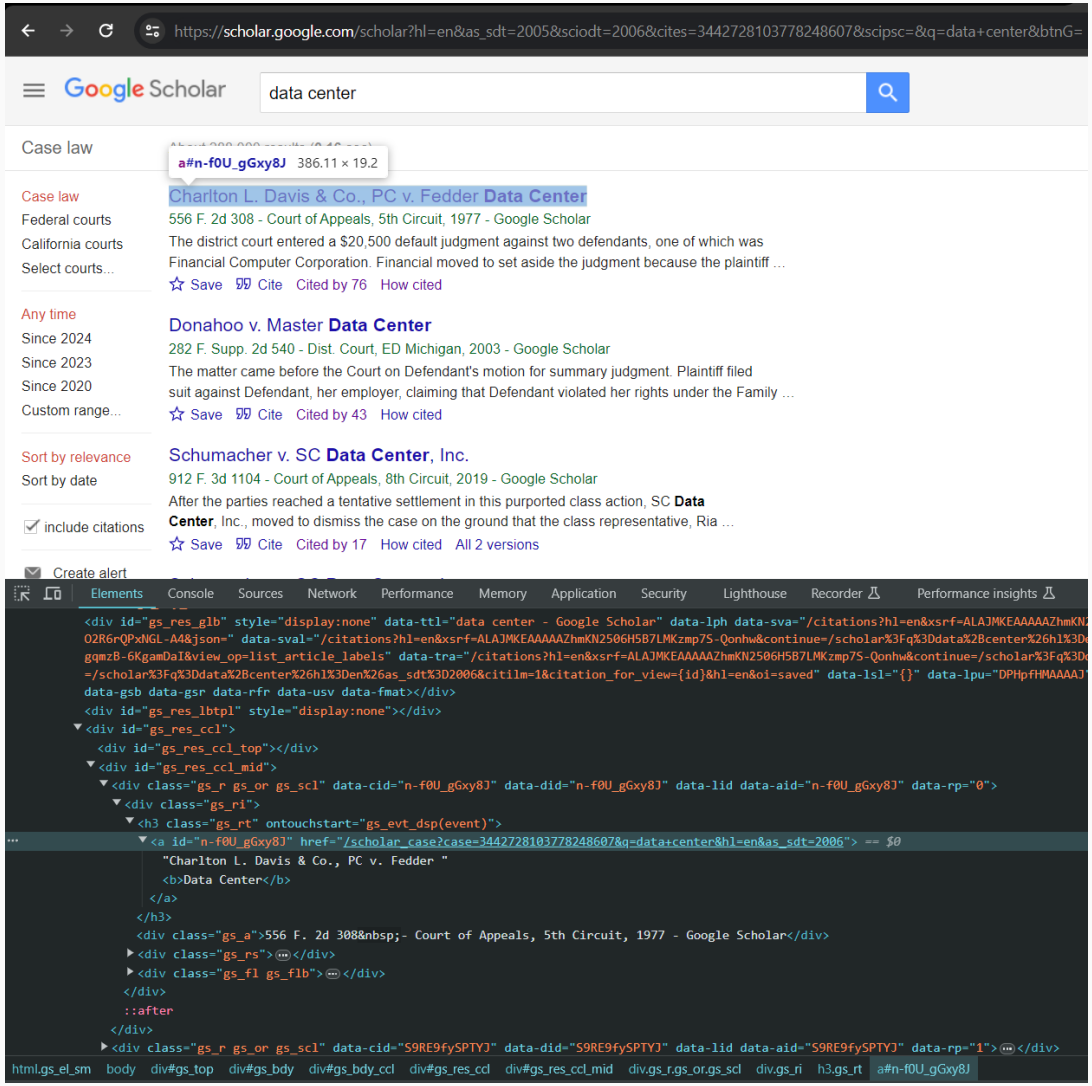


Figure 1: Google Scholar with developer tools highlighting the structure of a search results page.

Graph Visualization of Dataset

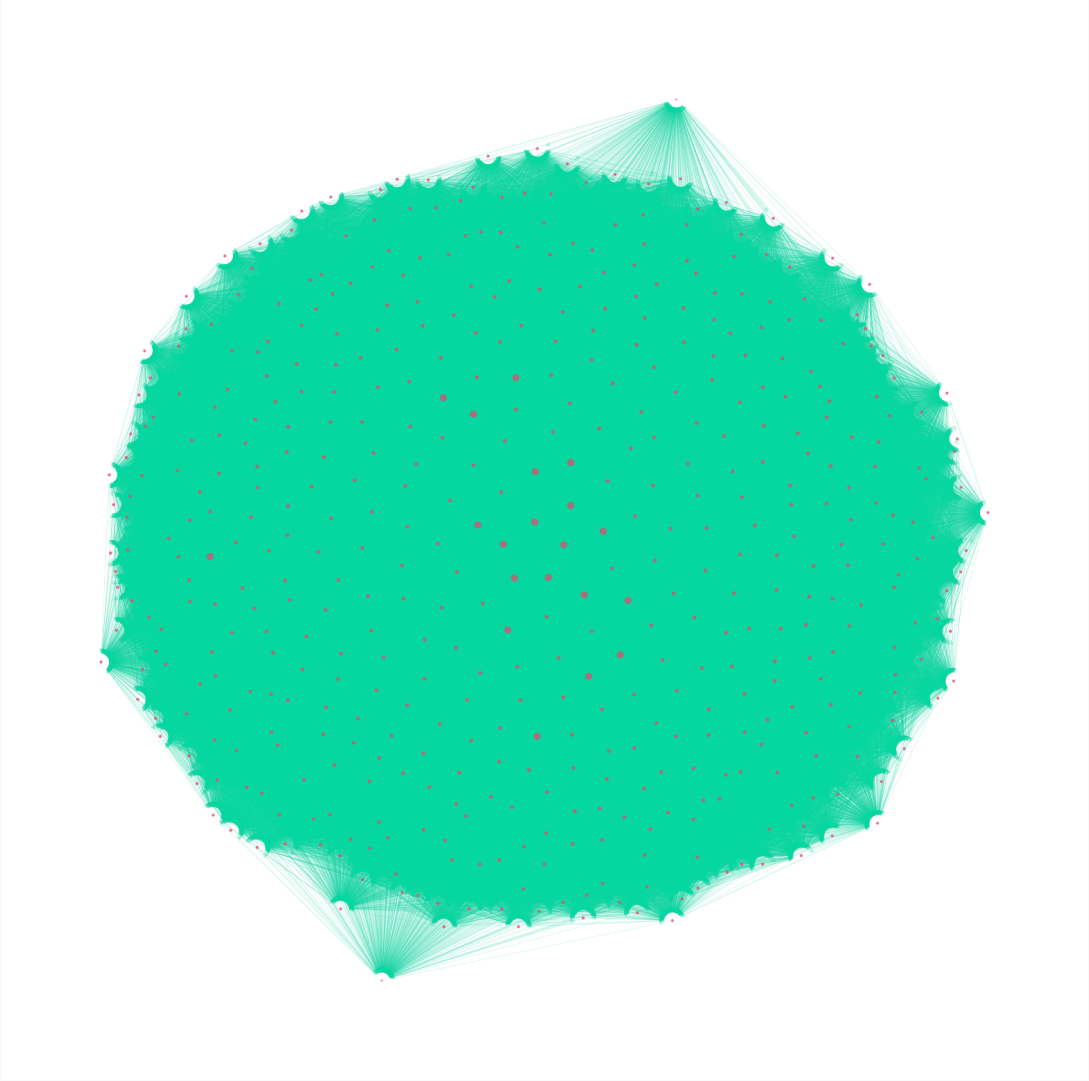


Figure 2: Visualization of the dataset with 500 nodes.

PageRank Comparison

case id	simple_pagerank	personalized_pagerank
qH2nNk7LUq0J	0.002318	0.009491
U6elx8mYkL4J	0.002318	0.002006
kWCO-ADQwtYJ	0.002318	0.009491
cR3BW4nVvwoJ	0.002318	0.002006
7oJcOnQz8QYJ	0.002318	0.002006
...	...	...
p8sBfgjmc44J	0.001367	0.001095
uSrZMaNnTXgJ	0.001293	0.000943
yoZxNbDWow8J	0.001356	0.000999
xX-bVtdoNGcJ	0.001467	0.001108
gDmV8jZ2LlkJ	0.000475	0.000221

Figure 3: Comparison of simple and personalized PageRank scores.

- **simple_pagerank**: Scores from the standard PageRank algorithm, based solely on network structure.
- **personalized_pagerank**: Scores adjusted by a personalization vector, which introduces bias towards certain nodes.

PageRank Comparison

case id	simple_pagerank	personalized_pagerank
qH2nNk7LUq0J	0.002318	0.009491
U6elx8mYkL4J	0.002318	0.002006
kWCO-ADQwtYJ	0.002318	0.009491
cR3BW4nVvwoJ	0.002318	0.002006
7oJcOnQz8QYJ	0.002318	0.002006
...	...	...
p8sBfgjmc44J	0.001367	0.001095
uSrZMaNnTXgJ	0.001293	0.000943
yoZxNbDWow8J	0.001356	0.000999
xX-bVtdoNGcJ	0.001467	0.001108
gDmV8jZ2LlkJ	0.000475	0.000221

Figure 4: Comparison of simple and personalized PageRank scores.

Key Observations:

- Variability between simple and personalized PageRank scores indicates the influence of personalization.
- Cases with higher personalized scores were likely prioritized in the personalization vector.
- The analysis reveals the impact of personalization on the authority of cases in a legal citation network.

PageRank Log Distributions

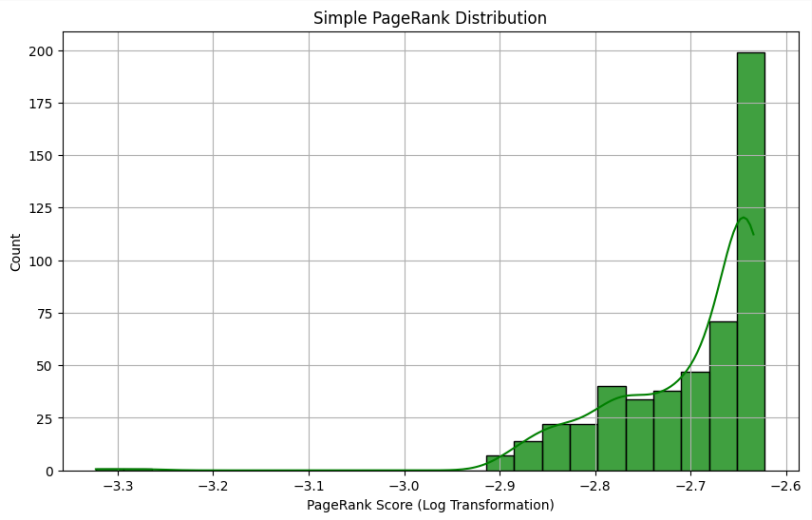


Figure 5: Personalized PageRank Distribution

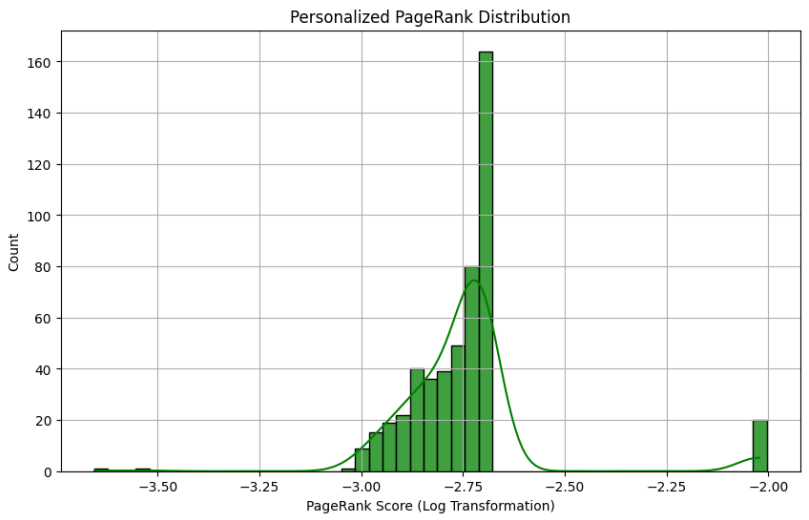


Figure 6: Simple PageRank Distribution

- The Simple PageRank shows a concentration of cases around a specific score range.
- The Personalized PageRank has a more uniform distribution across different scores, although there is still a peak
- This suggests that personalization introduces a bias towards certain nodes within the graph.

Future Improvements

Future Improvements

- Expanding the dataset to over 4.6 million cases.
- Weighting the data using relevance of each citation to the case.
- Refining geographical weighting and incorporating expert assessments in rankings.

Conclusion

So What is the top case?!

Lau's Corp., Inc. v. Haskins, 405 SE 2d 474 - Ga: Supreme Court 1991

Read How cited

405 S.E.2d 474 (1991)
261 Ga. 491

LAU'S CORPORATION, INC. d/b/a China King Restaurant
v.
Sarah HASKINS, et al.

No. S91G0720.

Supreme Court of Georgia.

June 27, 1991.
Reconsideration Denied July 24, 1991.

475 *475 Edwin A. Tate II, Jeanne F. Johnson, Bentley, Karesh, Seacrest, Labovitz & Campbell, Atlanta, for Lau's Corp., Inc.

Lester Z. Dozier, Dozier, Akin & Lee, Macon, for Haskins, et al.

CLARKE, Chief Justice.

Sarah and Louis Haskins were robbed by two men in the parking lot adjoining the China King Restaurant. Louis Haskins was hit in the head and Sarah Haskins' purse was snatched. The Haskinses brought an action against the China King Restaurant alleging that it failed to provide adequate warning or security for its patrons. The trial court granted summary judgment to the restaurant. The Court of Appeals reversed. *Haskins v. Lau's Corporation, Inc.*, 198 Ga.App. 470, 402 S.E.2d 58 (1991). We granted certiorari and reverse the Court of Appeals.

So what is the top case?!

This case, being the most significant in our dataset, revolves around a negligence claim against China King Restaurant by Sarah and Louis Haskins following a robbery and assault in the restaurant's parking lot. The legal question centered on the restaurant's duty of care and whether they provided adequate security.

Note: This case is key for its insights into a proprietor's duty of care and the limits of liability for unforeseen crimes, influencing the evaluation of similar cases.

Acknowledgements

I would like to express my deepest gratitude to those who have helped me throughout this project:

- My mentor, Arturo, for his invaluable guidance and support.
- The faculty and staff at Northeastern University for providing the necessary resources and for hosting the DRP

Questions?

References i

- Wikipedia: PageRank
- Tutorial on using Pagerank
- Time-Aware Weighted PageRank for Paper Ranking in Academic Graphs
- PageRank for bibliographic networks
- Bibliometric Measures in Citation-Tracking Databases
- Check out the code on github !

Thank You!