

Protecting PageRank: Helping Search Engines Maintain Result Integrity



Cordelia Ludden*, Helena Simson*, Sarah Radway^, Daniel Votipka* *Tufts University, ^Harvard University

BACKGROUND

PageRank represents the order in which Google Search results appear. It is determined by many features, such as the website's age and frequency of keyword occurrence.

Services have emerged that offer to manipulate PageRank to eliminate unfavorable results.

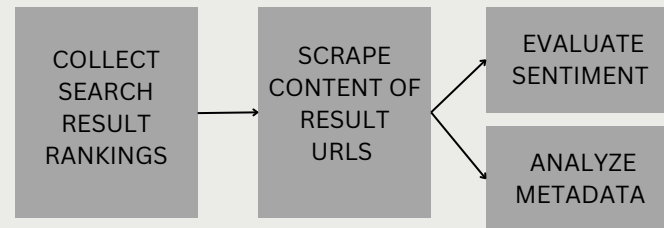


This damages result relevance and creates equity issues, as 9% of Google users reach the bottom of the first results page, and a mere 0.44% examine the second page.

RESEARCH QUESTION

What features are associated with artificial modifications of Google Page Rank?

METHODOLOGY



1. COLLECT SEARCH RESULT RANKINGS

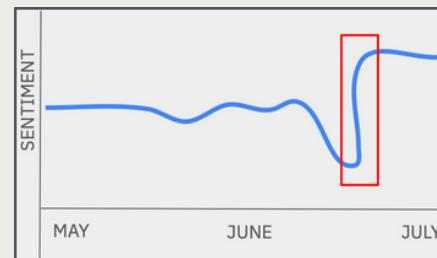
We search the names of the 1965 candidates running for congressional election in 2024 and the CEO, COO, and CFO of the Fortune 500 companies each day, using the Google Search API. We collect the first 100 results.

2. SCRAPE URL CONTENT

We scrape the site contents of all of ranked pages.

3. EVALUATE SENTIMENT

We perform relational sentiment analysis on the site content to identify dramatic changes in sentiment. We then manually review these periods of dramatic change, and attempt to link them to real-world events.



4. ANALYZE METADATA

We collect data surrounding data backlink quality and count, external link quality and count to support anomalous findings in (3).

INITIAL RESULTS

CHURN

RESULT A ↑ +1
RESULT B ↓ -1
RESULT C ■ NC

ENTRY/EXIT CHURN: the mean across all candidates is 1.52 (meaning ~1 entry enters /leaves the rankings throughout the week).

RANK CHURN: the mean of the summed rank churns across all candidates is 5.24, with the mean per site being 0.64.

Generally top result ordering stays the same. We can expect that anomalous events occur if an item changes five or more positions.

FUTURE WORK

We will be collecting six months worth of search results using the described methods, leading up to the 2024 election.

As this is a work in progress, we are actively seeking feedback on our methodology, and potential suggestions for how to improve it!